Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Académico Profesional de Estadística

Regresión logística y su aplicación en un caso de epidemiología

MONOGRAFÍA

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Zoila ORTIZ MIGUEL

ASESOR

Rosa Ysabel ADRIAZOLA CRUZ

Lima, Perú

2005

Referencia bibliográfica

Ortiz, Z. (2005). *Regresión logística y su aplicación en un caso de epidemiología.* Monografía para optar el título profesional de Licenciado en Estadística. Escuela Académico Profesional de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

REGRESIÓN LOGÍSTICA Y SU APLICACIÓN EN UN CASO DE EPIDEMIOLOGÍA

ZOILA ORTIZ MIGUEL

Monografía presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el Título Profesional de Licenciado en Estadística.

Aprobado por:

_____

Mg. Rosa Ysabel Adriazola Cruz

_____

Mg. María Estela Ponce Aruneri

Lima – Perú
Noviembre - 2005

ORTÍZ MIGUEL, ZOILA

Regresión Logística y su Aplicación en un caso de epidemiología, (Lima) 2005.

, 100 p., (UNMSM, Licenciado, Estadística, 2005).

Monografía, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas 1. Estadística I. UNMSM/FdeCM II.

A Dios, y a mis amados padres, por el inmenso amor que me tienen, por todo el apoyo que me dan, por haber hecho posible lograr mi profesión, y principalmente, por darle sentido a mi vida.

RESUMEN

REGRESIÓN LOGÍSTICA Y SU APLICACIÓN
EN UN CASO DE EPIDEMIOLOGÍA

<u>ZOILA ORTIZ MIGUEL</u>

NOVIEMBRE - 2005

ORIENTADOR : MG. ROSA YSABEL ADRIAZOLA CRUZ
TÍTULO OBTENIDO : LICENCIADO EN ESTADÍSTICA

- -

En este trabajo de investigación se presenta una aplicación de la Regresión Logística en el área de la epidemiología, con el objetivo de identificar los factores que determinan la presencia de asma en los escolares de 3 a 14 años de la ciudad de Moquegua.

En la parte inicial, se presenta la teoría de la Regresión Logística Binaria que fundamenta esta investigación, y finalmente, la aplicación e interpretación de resultados.

<u>PALABRAS CLAVES</u>: REGRESIÓN LOGÍSTICA
EPIDEMIOLOGÍA
ASMA
ODDS RATIO

ABSTRACT

LOGISTIC REGRESSION AND APLICATION IN A CASE OF EPIDEMIOLOGY

ZOILA ORTIZ MIGUEL

NOVEMBER 2005

ADVICER : MG. ROSA YSABEL ADRIAZOLA CRUZ

TITLE TO BE OBTAINED : LICENCIADA EN ESTADÍSTICA

In this research, we present an application of Logistic Regression in the epidemiology's area, with the purpose of to know the relevant factors in the presence of asthma in students between 3 to 14 years old of the Moquegua city.

In the first part, show the theory that supports this research, Binary Logistic Regression, and finally, one application and the interpretation of the results.

Key Words: Logistic Regression, Epidemiology, Asthma, Odds Ratio.

INDICE

INTRODUCCIÓN

La regresión logística es de lejos el modelo estadístico más popular, usado para analizar datos epidemiológicos. Su objetivo primordial es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías; las variables independientes pueden ser continuas, ó categóricas (nominales u ordinales) ó una mezcla de ambos tipos de variables.

La motivación principal para estudiar el análisis de regresión logística se debe a que frecuentemente en las investigaciones epidemiológicas, una pregunta típica de los investigadores es: ¿Cuál es la relación que existe entre una o más variables de exposición y una variable respuesta enfermedad?. A fin de evaluar la magnitud en la cual ese conjunto de factores de exposición están asociados con la enfermedad, es necesario controlar algunas variables confusoras, las cuales no son de interés principal.

En la presente monografía se abordará el modelo de regresión logística para respuesta dicotómico, a través de una aplicación en el área de la epidemiología; se estudiará la influencia de ciertos factores en la presencia de la enfermedad respiratoria asma en escolares de la ciudad de Moquegua. El software estadístico utilizado es el SPSS, debido a su uso generalizado en los profesionales tanto de la salud como de otras especialidades.

En este análisis de Regresión Logística se construye una ecuación de regresión para predecir la presencia de asma a partir de una combinación lineal de variables. La ecuación dará el riesgo de contraer asma con una suma ponderada de los factores.

$$\hat{g}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$$

Las probabilidades se encuentran entre 0 y 1, las que se transforman a una escala de valores de $\hat{g}(X)$, y se les denomina transformación logística:

Sí $\hat{Y} = P\left(E / \hat{g}(X) \right) = \dfrac{1}{1 + e^{-\hat{g}(X)}}$, donde $E =$ Presencia de asma

$\hat{Y}$ es la probabilidad de contraer asma para un valor conocido de $\hat{g}(X)$.

Entre las ventajas de este modelo tenemos: las estimaciones se encuentran en el rango de 0 a 1 (probabilidad de contraer la enfermedad), el modelo describe los efectos combinados de diferentes factores de riesgo, en la posibilidad de contraer la enfermedad, la interpretación del modelo también es posible a través de los odds ratio (funciones de parámetros del modelo).

Espero que esta aplicación logre mostrar la utilidad del modelo de regresión logística en el campo de la epidemiología, así como el procedimiento de análisis que debe seguirse.

CAPÍTULO I

REGRESIÓN LOGÍSTICA BINARIA

1.1 FORMULACIÓN DEL MODELO

Sin duda alguna la regresión logística se ha constituido en uno de los principales métodos de análisis estadístico de datos. Asimismo cuando la respuesta de interés no es originalmente de tipo binario, algunos investigadores dicotomizan la respuesta de modo que la probabilidad de éxito pueda ser modelada a través de la regresión logística.

Cuando la variable dependiente es una variable categórica binaria, la metodología de la regresión lineal no es aplicable ya que ahora la variable respuesta sólo presenta dos valores (nos centraremos en el caso dicotómico), como por ejemplo puede ser presencia / ausencia de hipertensión. Si clasificamos el valor de la variable respuesta como 0 cuando no se presenta el suceso (ausencia de hipertensión) y con el valor 1 cuando sí está presente (paciente hipertenso), y buscamos cuantificar la posible relación entre la presencia de hipertensión y, por ejemplo, la cantidad media de sal consumida al día como posible factor de riesgo, podríamos caer en la tentación de utilizar una regresión lineal:

$$\text{Hipertensión} = \beta_0 + \beta_1 X_1 + \varepsilon, \qquad \text{donde } X_1 = \text{Consumo de sal}$$

y estimar, a partir de nuestros datos, por el procedimiento habitual de mínimos cuadrados, los coeficientes β_0 y β_1 de la ecuación. Sin embargo, y aunque esto es posible matemáticamente, nos conduce a la obtención de resultados absurdos, ya que cuando se calcule la función obtenida para diferentes valores de consumo de sal se obtendrá resultados que, en general, serán diferentes de 0 y 1, los únicos

realmente posibles en este caso, ya que esa restricción no se impone en la regresión lineal, en la que la respuesta puede en principio tomar cualquier valor.

Problemas que pueden presentarse al utilizar un modelo de regresión lineal cuando la variable respuesta es dicotómica:

1. **Los errores no tienen distribución normal.**

Cada error $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$, puede asumir uno de los dos siguientes valores:

$$Y_i = 1 \Rightarrow \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$
$$Y_i = 0 \Rightarrow \varepsilon_i = -\beta_0 - \beta_1 X_i$$

Es decir, los errores se distribuyen en forma binomial

2. **Varianza heterogénea.**

La varianza de Y_i para el modelo de regresión lineal simple es:

$$\sigma^2(Y_i) = E[(Y_i - E(Y_i))^2] = (1 - \Pi(X_i))^2 \Pi(X_i) + (0 - \Pi(X_i))^2 (1 - \Pi(X_i))$$
$$\sigma^2(Y_i) = \Pi(X_i)(1 - \Pi(X_i)) = E(Y_i)(1 - E(Y_i))$$

$$\varepsilon_i = Y_i - \Pi(X_i), \quad (\Pi(X_i) \text{ constante})$$

Como

$$\sigma^2(\varepsilon_i) = \Pi(X_i)(1 - \Pi(X_i)) = (\beta_0 + \beta_1 X_i) + (1 - \beta_0 - \beta_1 X_i)$$

Tenemos

$$\uparrow$$
Depende de X_i

Es decir, los errores no cumplen el supuesto de homocedasticidad (varianza no es constante)

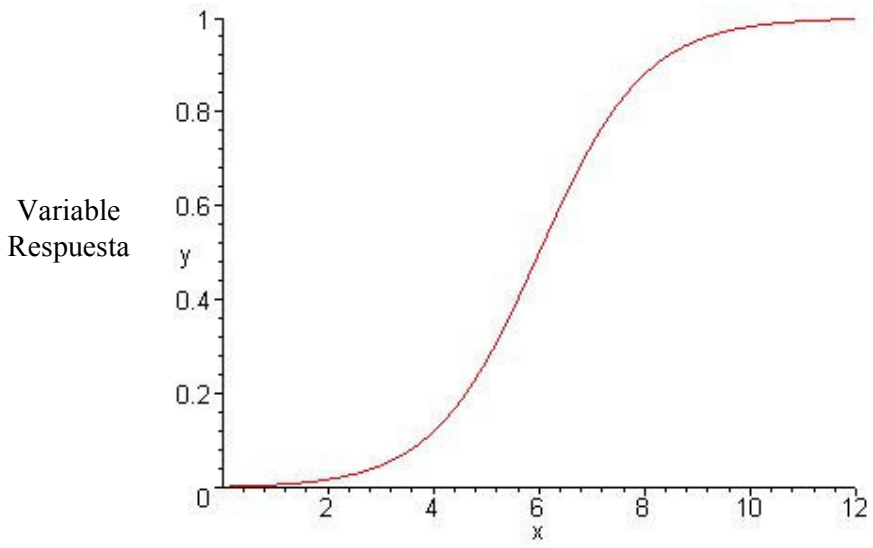
3. Restricción de la variable respuesta.

Como la función respuesta representa probabilidades cuando la variable respuesta es binaria, entonces:

$$\boxed{0 \le E(Y) = \Pi(X_i) \le 1}$$

La restricción en la respuesta media de presentar valores en el intervalo [0,1] es inapropiada para una función de respuesta lineal, debido a que el extremo derecho de la ecuación de regresión lineal no necesariamente producirá valores en ese intervalo.

Para resolver este último problema, utilizamos la función logística que da origen al modelo de regresión logística y que se muestra a continuación.



Una propiedad interesante de la función logística es que puede ser linealizada. Sea $E(Y)=\Pi(X)$, debido a que la respuesta media es la probabilidad cuando la variable respuesta es binaria. Realizando la transformación:

$$\Pi'(X)=\log_e\left(\frac{\Pi(X)}{1-\Pi(X)}\right)$$

Obtenemos

$$\Pi'(X)=\beta_0+\beta_1 X \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Esta transformación es llamada la transformación Logit de la probabilidad $\Pi(X)$, la razón $\Pi(X)/(1-\Pi(X))$ de la transformación Logit es llamado el Odds (chance). La función respuesta transformada (1), se denomina función respuesta logit, y $\Pi(X)$ se denomina la respuesta media logit.

Consideremos el modelo de regresión logístico lineal simple en el que $\Pi(X)$ es la probabilidad de "éxito" dado el valor X de una variable explicativa cualquiera, el cual se define como:

$$Log\left[\frac{\Pi(X)}{1-\Pi(X)}\right] = \beta_0 + \beta_1 X$$

Donde: β_0 y β_1, son los parámetros desconocidos. Este modelo podría ser aplicado al ejemplo anterior, asociación entre la presencia de hipertensión y la cantidad media de sal consumida. Serían entonces muestreados, independientemente, n_1 individuos con presencia del factor (X=1) y n_2 individuos con ausencia del factor (X=0), $\Pi(X)$sería la probabilidad de que un paciente padezca hipertensión. De

esta forma, la chance de desarrollar hipertensión para un individuo con presencia del factor esta dado por:

$$\frac{\Pi(1)}{1-\Pi(1)} = e^{\beta_0+\beta_1}$$

en cambio la chance de desarrollar hipertensión para un individuo con ausencia del factor es simplemente

$$\frac{\Pi(0)}{1-\Pi(0)} = e^{\beta_0}$$

Luego, la razón de chances esta dada por

$$\Psi = \frac{\Pi(1)[1-\Pi(0)]}{\Pi(0)[1-\Pi(1)]} = e^{\beta}$$

Como el número de factores puede ser más de uno, el modelo general de regresión logística binaria se expresa como:

$$Log\left[\frac{\Pi(X)}{1-\Pi(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Donde $X=(1, X_1, X_2, ..., X_p)^T$ contiene los valores observados de p variables explicativas, y $\beta_0, \beta_1\ \beta_2 \ldots \beta_p$ son los parámetros del modelo; así para el ejemplo anterior de hipertensión, en el extremo derecho de la ecuación podríamos tener:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

donde: X_1 = Consumo de sal, X_2= Edad, X_3 = Fumador

LAS VARIABLES CUALITATIVAS EN EL MODELO LOGÍSTICO

Puesto que la metodología empleada para la estimación del modelo logístico se basa en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, es incorrecto que en él intervengan variables cualitativas, ya sean nominales u ordinales.

La asignación de un número a cada categoría no resuelve el problema ya que si tenemos, por ejemplo, la variable ejercicio físico con tres posibles respuestas: sedentario, realiza ejercicio esporádicamente, realiza ejercicio frecuentemente, y le asignamos los valores 0, 1, 2, significa a efectos del modelo, que efectuar ejercicio físico frecuentemente es dos veces mayor que solo hacerlo esporádicamente, lo cual no tienen ningún sentido. Más absurdo sería si se trata, a diferencia de ésta, de una variable nominal, sin ninguna relación de orden entre las respuestas, como puede ser el estado civil.

La solución a este problema es crear tantas variables dicotómicas como número de categorías - 1. Estas nuevas variables, artificialmente creadas, reciben en la literatura anglosajona el nombre de "*dummy*", traduciéndose en español con diferentes denominaciones como pueden ser variables internas, indicadoras, o variables diseño.

Estas variables se pueden construir de diversas formas y según la forma utilizada, cambia la interpretación de sus coeficientes β_i que, en general, cuantifican el efecto de un valor de dichas variables con respecto a un valor de referencia

Así por ejemplo (TABLA 1.1), si la variable en cuestión recoge datos de tabaquismo con las siguientes respuestas: Nunca fumó, Ex-fumador, Actualmente fuma menos de 10 cigarrillos diarios, Actualmente fuma 10 o más cigarrillos diarios, tenemos 4 posibles respuestas por lo que construiremos 3 variables internas dicotómicas (valores 0,1), existiendo diferentes posibilidades de codificación, que conducen a diferentes interpretaciones, siendo la más habitual la codificación indicador.

TABLA 1.1 CODIFICACIÓN INDICADOR

Categorías de la variable Tabaquismo	X_1	X_2	X_3
Nunca fumó	0	0	0
Ex- fumador	1	0	0
Menos de 10 cigarrillos diarios	0	1	0
10 o más cigarrillos diarios	0	0	1

En este tipo de codificación el coeficiente de la ecuación de regresión para cada variable diseño (siempre transformado con la función exponencial), se corresponde al odds ratio de esa categoría con respecto al nivel de referencia (Nunca fumó), en nuestro ejemplo cuantifica cómo cambia el riesgo respecto a no haber fumado nunca.

Además existen otras posibilidades de codificación entre las que tenemos diferencia, desviación, etc.

INTERACCIÓN Y CONFUSIÓN

El empleo de análisis de regresión es útil para dos objetivos:

- Estimar la relación entre dos variables teniendo en cuenta la presencia de otros factores

- Construir un modelo que permita predecir el valor de la variable dependiente (en regresión logística la probabilidad del suceso) para unos valores determinados de un conjunto de variables pronóstico

Cuando el objetivo es estimar la relación o asociación entre dos variables, los modelos de regresión permiten considerar que puede haber otros factores que modifiquen esa relación.

Así, por ejemplo, si estamos estudiando la posible relación, como factor de riesgo, entre el síndrome de apnea nocturna y la probabilidad de padecer hipertensión, dicha relación puede ser diferente si se tiene en cuenta otras variables como pueden ser la edad, el sexo o el índice de masa corporal. Por ello, en un modelo de regresión logística podrían ser incluidas otras variables independientes, además del apnea nocturna. En la ecuación obtenida al considerar como variables independientes *Apnea, Edad, Sexo, IMC,* el *exp(coeficiente de la ecuación para Apnea)* determina el odds ratio debido al apnea, ajustado o controlado para el resto de los factores.

A las variables que, además del factor de interés (en el ejemplo *Edad, Sexo, IMC*), se introducen en el modelo, se las denomina de diferentes formas: variables control, variables extrañas, covariantes, o factores de confusión.

Cuando la relación entre el factor en estudio y la variable dependiente se modifica según el valor de una tercera estamos hablando de interacción. Así en nuestro ejemplo, supongamos que la probabilidad de padecer *Hipertensión* cuando se tiene síndrome de apnea aumenta con la edad. En este caso decimos que existe interacción entre las variables *Edad* y *Apnea*.

Si nos fijamos sólo en el exponente del modelo logístico, en el caso de no considerar interacción éste será:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2$$

donde: X_1 = Apnea, X_2 = Edad

Si consideramos la presencia de interacción entre Apnea y Edad, el modelo cambia a:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Si la variable *Apnea* es dicotómica (valores 0 y 1) la relación entre *Hipertensión* y *Apnea* vendrá cuantificada por β_1 en el primer modelo, mientras que en el segundo será:

$$(\beta_1 + \beta_3 X_2) X_1$$

es decir, que ahora esa relación se modifica en función del valor de la *EDAD*.

Una de las primeras consideraciones que hay que tener en cuenta es que la relación entre la variable independiente y la probabilidad de éxito no cambie de sentido, ya que en ese caso no nos sirve el modelo logístico. Esto es algo que habitualmente no ocurre en los estudios clínicos, pero por ello es más fácil pasarlo por alto cuando se presenta.

Un ejemplo muy claro de esa situación se da si estamos evaluando la probabilidad de nacimiento de un niño con bajo peso (de riesgo) en función de la edad de la madre. Hasta una edad esa probabilidad puede aumentar a medida que la edad de la madre disminuye (madres muy jóvenes) y a partir de una edad (madres muy mayores) la probabilidad puede aumentar a medida que lo hace la edad de la madre. En este caso el modelo logístico no sería adecuado.

SELECCIÓN DE VARIABLES

Un paso importante en la construcción de un modelo de regresión es el de, la elección de variables a incluir. Los mecanismos para la selección de variables no son fáciles de especificar ya que dependen en gran medida del tipo de modelo (predictivo o explicativo), del contexto de utilización y de las propias características del proceso analizado. Quizás la única norma clara es que ante dos posibles modelos, similares en otros aspectos, preferiremos el que sea más sencillo y que menos suposiciones necesite para su construcción (es lo que se denomina principio de parsimonia).

Para poder decidir entre utilizar un modelo con unas determinadas variables o con otras será preciso disponer de una medida de comparación entre modelos.

En la regresión lineal se utiliza para comparar dos modelos la ***F* parcial**, que en el caso de que se contrasten dos modelos que difieren en una sola variable es idéntico a utilizar el valor de la ***t*** para el coeficiente de regresión de la nueva variable.

En la regresión logística, y en general en cualquier modelo de regresión cuyos coeficientes se estimen por el método de máxima verosimilitud, se utiliza el **cociente de verosimilitud**, que es una medida, a partir de los datos de nuestra muestra, de cuánto más probable (verosímil) es un modelo frente al otro. Este parámetro se distribuye según una Chi-cuadrado con grados de libertad igual a la diferencia entre el número de variables de los dos modelos. Si no es suficientemente grande decimos que no hay evidencia para pensar que un modelo es mejor que el otro y por tanto nos quedaremos con el más sencillo.

El conjunto de variables que finalmente quede incluido en la ecuación de regresión puede depender del camino seguido a la hora de seleccionarlas, salvo en el caso de que se evalúen todos los modelos de regresión posibles que obviamente sólo tiene una conclusión.

Cualquiera que sea el método que se piense utilizar para la selección de variables éste debe comenzar con un cuidadoso análisis univariable de la posible relación entre la variable dependiente y cada uno de los factores estudiados.

Una vez definido el conjunto de variables a incluir en el modelo logístico, resta saber cual es la mejor manera de encontrar un modelo reducido que incluya apenas las covariables e interacciones más importantes para explicar la probabilidad de éxito $\Pi(X)$. Este problema podría ser resuelto por los métodos usuales de selección de modelos.

La definición de mejor modelo depende del tipo y el objetivo del estudio. En un modelo con finalidad predictiva se considerará como mejor modelo aquel que produce predicciones más fiables, mientras que en un modelo que pretende estimar la relación entre dos variables (corrigiendo el efecto de otras), se considerará mejor aquel con el que se consigue una estimación más precisa del coeficiente de la variable de interés. Esto se olvida a menudo y sin embargo conduce a estrategias de modelado completamente diferentes. Así en el segundo caso una covariable con coeficiente estadísticamente significativo pero cuya inclusión en la ecuación no modifica el valor del coeficiente de la variable de interés, será excluído de la ecuación, ya que no se trata de un factor de confusión: la relación entre la variable de interés y la probabilidad no se modifica si se tiene en cuenta esa variable. Sin embargo si lo que se busca es un modelo predictivo sí que se incluirá en la ecuación pues ahora lo que buscamos es predicciones más fiables.

Lo primero que habrá que plantear es el modelo máximo, o lo que es lo mismo el número máximo de variables independientes que pueden ser incluidas en la ecuación, considerando también las interacciones si fuera conveniente.

Aunque existen diferentes procedimientos para escoger el modelo sólo hay tres mecanismos básicos para ello: empezar con una sola variable independiente e ir añadiendo nuevas variables según un criterio prefijado (procedimiento hacia delante "Forward"), o bien empezar con el modelo máximo e ir eliminando de él variables según un criterio prefijado (procedimiento hacia atrás "Backward"). El tercer y más aplicado método, denominado en la literatura "S*tepwise*" , combina los dos anteriores y en cada nuevo paso, cuando se incluye una nueva variable, además se reconsidera el mantener las que ya se había añadido previamente, es decir que no sólo puede entrar una nueva variable en cada paso sino que puede salir alguna de las que ya estaban en la ecuación; este proceso finaliza cuando ninguna variable de las que no están en la ecuación cumple la condición para entrar y de las incorporadas a la ecuación ninguna cumple la condición para salir.

En el caso de la regresión logística el criterio para decidir en cada paso si escogemos un nuevo modelo frente al actual viene dado por el logaritmo del cociente de verosimilitudes de los modelos como lo mencionamos anteriormente.

La función de verosimilitud de un modelo es una medida de cuán compatible es éste con los datos realmente observados. Si al añadir una nueva variable al modelo

no mejora la verosimilitud de forma apreciable, en sentido estadístico, ésta variable no se incluye en la ecuación.

Para evaluar la significación estadística de una variable concreta dentro del modelo, nos fijaremos en el valor de Chi-cuadrado del estadístico de Wald correspondiente al coeficiente de la variable y en su nivel de probabilidad.

La interpretación de los parámetros de regresión es crucial en el modelo logístico, lo cual implica que una forma puramente mecánica de selección de variables puede llevar a un modelo sin sentido y de difícil interpretación.

Muchas veces, las variables consideradas biológicamente importantes no deben ser dejadas de lado por su falta de significancia estadística, la selección de un modelo logístico debe ser un proceso conjunto de selección estadística de modelos y buen sentido.

LOS COEFICIENTES DEL MODELO LOGÍSTICO COMO CUANTIFICADORES DE RIESGO

Una de las características que hacen tan interesante la regresión logística es la relación que éstos guardan con un parámetro de cuantificación de riesgo conocido en la literatura como "odds ratio". El odds asociado a un suceso es el cociente entre la probabilidad de que ocurra frente a la probabilidad de que no ocurra:

$$\text{Odds} = \frac{\Pi(X)}{1-\Pi(X)}$$

siendo $\Pi(X)$ la probabilidad de éxito. Así, por ejemplo, podemos calcular el odds de presencia de hipertensión cuando el consumo diario de sal es igual o superior a una cierta cantidad, que en realidad determina cuántas veces es más probable que haya hipertensión a que no la haya en esa situación. Igualmente podríamos calcular el odds de presencia de hipertensión cuando el consumo de sal es inferior a esa cantidad. Si dividimos el primer odds entre el segundo, hemos calculado un cociente de odds, esto es un odds ratio, que de alguna manera cuantifica cuánto más probable es la aparición de hipertensión cuando se consume mucha sal (primer odds) respecto a cuando se consume poca. La noción que se está midiendo es parecida a la que encontramos en lo que se denomina riesgo relativo que corresponde al cociente de la probabilidad de que aparezca un suceso (hipertensión) cuando está presente el factor (consumo elevado de sal) respecto a cuando no lo está. De hecho cuando la prevalencia del suceso es baja (< 20 %) el valor del odds ratio y el riesgo relativo es muy parecido, pero no es así cuando el suceso es bastante común, hecho que a menudo se ignora.

Si en la ecuación de regresión tenemos un factor dicotómico, como puede ser por ejemplo si el sujeto es no fumador, el coeficiente β de la ecuación para ese factor está directamente relacionado con el odds ratio (OR) de ser fumador respecto a no serlo

$$\text{OR} = \exp(\beta)$$

es decir que $\exp(\beta)$ es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes. Cuando la variable es numérica,

como puede ser por ejemplo la edad, o el índice de masa corporal, es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro, permaneciendo constantes el resto de variables. Así el odds ratio que supone pasar de la edad X_1 a la edad X_2, siendo β el coeficiente correspondiente a la edad en el modelo logístico es:

$$\text{OR} = \exp[\beta(X_2 - X_1)]$$

Nótese que se trata de un modelo en el que el aumento o disminución del riesgo al pasar de un valor a otro del factor es proporcional al cambio, es decir a la diferencia entre los dos valores, pero no al punto de partida, quiere esto decir que el cambio en el riesgo, con el modelo logístico, es el mismo cuando pasamos de 40 a 50 años que cuando pasamos de 80 a 90. Cuando el coeficiente β de la variable es positivo obtendremos un odds ratio mayor que 1 y corresponde por tanto a un factor de riesgo. Por el contrario, si β es negativo, el odds ratio será menor que 1 y se trata de un factor de protección.

1.2 ESTIMACIÓN DE PARÁMETROS

Supongamos que tenemos una muestra de n observaciones independientes del par (X_i,Y_i), i = 1,2,..,n; donde Y_i denota el valor de una variable respuesta dicotómica y X_i es el valor de la variable independiente para el i-ésimo sujeto.

La variable respuesta ha sido codificada como 0 y 1, representando la ausencia ó presencia de la característica respectivamente.

Para ajustar el modelo de regresión logístico

$$\Pi(X) = \frac{e^{g(X)}}{1+e^{g(X)}}, \qquad g(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta p X_p$$

a un conjunto de datos, requiere que estimemos los valores de β'= (β_0 , β_1 ,..., βp), los parámetros desconocidos.

En regresión lineal, el método más frecuentemente usado para estimar parámetros desconocidos es el de mínimos cuadrados. En ese método se eligen aquellos valores de β que minimizan la suma de desviaciones al cuadrado de los valores observados y de los valores predichos basados en el modelo.

Bajo las suposiciones usuales para regresión lineal, el método de mínimos cuadrados produce estimadores con un número deseable de propiedades estadísticas. Desafortunadamente cuando el método de mínimos cuadrados se aplica a un modelo con respuesta dicotómica los estimadores no tienen estas mismas propiedades.

El método general de estimación que conduce a la función mínimos cuadrados bajo el modelo de regresión se llama máxima verosimilitud. En sentido general el método de máxima verosimilitud produce valores para los parámetros desconocidos, los cuales maximizan la probabilidad de obtener el conjunto de datos observados. Para aplicar este método debemos primero construir una función, llamada la función de verosimilitud, la cual expresa la probabilidad de los datos observados como una función de los parámetros desconocidos.

Los estimadores máximo verosímiles de estos parámetros son elegidos de aquellos valores que maximizan esta función. Así los estimadores resultantes son aquellos que concuerdan más cercanamente con los datos observados. A continuación se explica como encontrar estos valores del modelo de regresión logística.

Si Y es codificada como 0 ó 1, entonces $\Pi(X)$ provee (para un valor arbitrario de β, el vector de parámetros) la probabilidad condicional de Y=1 dado X, denotado como P(Y=1/X), de este modo, $1-\Pi(X)$ es la probabilidad condicional de Y=0 dado X, P(Y=0/X). Así, para estos pares (X_i,Y_i) donde Y_i =1, la contribución a la función verosimilitud es $\Pi(X_i)$, y para aquellos pares donde Y_i =0, la contribución a la función de verosimilitud es $1-\Pi(X_i)$, donde la cantidad $\Pi(X_i)$ denota el valor de $\Pi(X)$ calculado en X_i. Una forma conveniente de expresar la contribución a la función de verosimilitud del par (X_i,Y_i) es a través de la expresión:

$$\Pi(X_i)^{Y_i}\left[1-\Pi(X_i)\right]^{1-Y_i}$$

Ya que las observaciones son independientes, la función de verosimilitud se obtiene como el producto de los términos dados en la expresión anterior como sigue:

$$\lambda(\beta)=\prod_{i=1}^{n}\Pi(X_i)^{Y_i}\left[1-\Pi(X_i)\right]^{1-Y_i}$$

El principio de máxima verosimilitud establece que usamos como nuestra estimación de β, el valor que maximiza la expresión en la ecuación anterior. Sin embargo, es más fácil matemáticamente trabajar con el logaritmo de dicha ecuación. Así, el logaritmo de la verosimilitud se define como:

$$L(\beta) = \ln[\lambda(\beta)] = \sum_{i=1}^{n} \{Y_i \ln[\Pi(X_i)] + (1 - Y_i)\ln[1 - \Pi(X_i)]\}$$

Para encontrar el valor de β que maximiza L(β), diferenciamos L(β) con respecto a los p+1 coeficientes e igualamos las expresiones resultantes a 0. estas ecuaciones, conocidas como las ecuaciones de verosimilitud son:

$$\sum [Y_i - \Pi(X_i)] = 0 \qquad (1.2.1)$$

$$\sum X_{ij}[Y_i - \Pi(X_i)] = 0 \qquad (1.2.2)$$

En regresión lineal, las ecuaciones de verosimilitud que se obtienen diferenciando la suma de desviaciones al cuadrado con respecto a β son lineales en los parámetros desconocidos y por ello fácilmente resueltos.

Para regresión logística, las ecuaciones de verosimilitud no son lineales en los parámetros y por ello requieren métodos especiales para su solución (Método iterativo de Newton Raphson). Estos métodos son iterativos por naturaleza y están disponibles en la mayoría de softwares de regresión logística.

El valor de β obtenido en la solución de las ecuaciones de verosimilitud se llama la estimación máximo verosímil y se denota como $\hat{\beta}$.

El método de estimación de varianzas y covarianzas de los coeficientes estimados sigue una bien desarrollada teoría de estimación de máxima verosimilitud. Esta

teoría establece que los estimadores se obtienen de la matriz de segundas derivadas parciales de la función logaritmo de la verosimilitud. Estas derivadas parciales tienen la siguiente forma general:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} X_{ij}^2 \Pi_i (1 - \Pi_i)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = \sum_{i=1}^{n} X_{ij} X_{il} \Pi_i (1 - \Pi_i)$$

Para j,l = 0,1,2,…,p donde $\Pi_i = \Pi(X_i)$. Sea la matriz (p+1)x(p+1) que contienen los términos negativos dados en la ecuación (1.2.1) y (1.2.2) denotados como $I(\beta)$. Esta matriz es llamada la matriz información observada. Las varianzas y covarianzas de los coeficientes estimados se obtienen del inverso de esta matriz, la cual denotamos como Var $(\beta) = I^{-1}(\beta)$.

1.3 INTERPRETACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICO

Supongamos que un modelo de regresión logístico ha sido ajustado, que las variables son significativas en el sentido clínico o estadístico y que el modelo ajusta bajo alguna medida estadística de bondad de ajuste.

La pregunta que debemos responder es: ¿Qué nos dicen los coeficientes estimados en el modelo acerca de las preguntas de investigación que motivaron el estudio? .

Para la mayoría de modelos esto involucra a los coeficientes estimados de las variables independientes en el modelo, es decir, la pendiente (razón de cambio) de una función de la variable dependiente por unidad de cambio en la variable independiente.

La interpretación implica dos propósitos: determinar la relación funcional entre la variable dependiente y la variable independiente, y definir apropiadamente la unidad de cambio para la variable independiente.

El primer paso es determinar que función de la variable dependiente produce una función lineal de las variables independientes, esta es llamada la función enlace. En el caso del modelo de regresión lineal, es la función identidad, debido a que la variable dependiente, por definición, es lineal en los parámetros. En el modelo de regresión logístico la función enlace es la transformación logit

$$g(X) = \ln\left\{\frac{\Pi(X)}{[1-\Pi(X)]}\right\} = \beta_0 + \beta_1 X .$$

Para el modelo de regresión lineal el coeficiente pendiente, β_1, es igual a la diferencia entre el valor de la variable dependiente en X+1 y el valor de la variable dependiente en X, para cualquier valor de X. Por ejemplo, si $Y(X) = \beta_0 + \beta_1 X$, entonces $\beta_1 = Y(X+1) - Y(X)$.

En este caso, la interpretación de los coeficientes es relativamente directa, este expresa el cambio resultante en la escala de medida de la variable dependiente por unidad de cambio en la variable independiente. Por ejemplo, si en una regresión del peso en función de la altura de adolescentes masculinos, la pendiente es 15,

podemos concluir que un incremento de un centímetro en la altura se asocia con un incremento de 15 gramos en el peso.

En el modelo de regresión logístico, el coeficiente pendiente representa el cambio en el logit correspondiente a una unidad de cambio en la variable independiente, es decir, $\beta_1 = g(X+1) - g(X)$.

A continuación consideramos la interpretación de los coeficientes para un modelo de regresión logístico univariable para cada una de las posibles escalas de medida de la variable independiente.

VARIABLE INDEPENDIENTE DICOTÓMICA

Sea la variable independiente X de escala nominal y dicotómica, esta es codificada como 0 ó 1. La diferencia en el logit para un sujeto con X=1 y X=0 es:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

Los posibles valores de las probabilidades logísticas pueden mostrarse convenientemente en una tabla 2×2 como la que se muestra en la tabla 1.3.1.

El odds cuando el evento de interés está presente entre los individuos con X=1 se define como $\frac{\Pi(1)}{[1-\Pi(1)]}$. Similarmente, el odds cuando el evento de interés está presente entre los individuos con X=0 se define como $\frac{\Pi(0)}{[1-\Pi(0)]}$. El odds ratio, denotado como OR, se define como la razón del odds para X=1 al odds para X=0, y está dado por la ecuación

$$OR = \frac{\Pi(1)/[1-\Pi(1)]}{\Pi(0)[1-\Pi(0)]} \qquad (1.3.1)$$

Sustituyendo las expresiones para el modelo de regresión logística obtenemos la tabla siguiente:

TABLA 1.3.1 VALORES DEL MODELO DE REGRESIÓN LOGÍSTICA CUANDO LA VARIABLE INDEPENDIENTE ES DICOTÓMICA

Variable Respuesta (Y)	Variable Independiente (X)	
	X=1	X=0
Y=1	$\Pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\Pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y=0	$1-\Pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\Pi(0) = \frac{1}{1+e^{\beta_0}}$

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)}$$

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Por lo tanto, para la regresión logística con una variable independiente dicotómica codificada como 1 y 0, la relación entre el odds ratio y el coeficiente de regresión es:

$$OR = e^{\beta_1}$$

Esta relación simple entre el coeficiente y el OR es la razón fundamental de porque la regresión logística se ha convertido en una herramienta de investigación analítica potente.

El OR es una medida de asociación que encontró amplio uso especialmente en epidemiología, ya que este aproxima cuanto más probable (ó improbable) es que la respuesta este presente entre aquellos con X=1 en comparación con aquellos con X=0.

La interpretación dada por el OR se basa en que muchas veces este aproxima una cantidad llamada el riesgo relativo, este parámetro es igual a la razón $\Pi(1)/\Pi(0)$.
De la fórmula (1.3.1) se deduce que el OR aproxima el riesgo relativo si $(1-\Pi(0))/(1-\Pi(1)) \approx 1$, esto se mantiene cuando $\Pi(X)$ es pequeño para X=1 y X=0.

El OR es usualmente el parámetro de interés en la regresión logística debido a su fácil interpretación. Sin embargo, su estimación, $\hat{OR}$, tiende a tener una distribución que es asimétrica. La asimetría de la distribución muestral de $\hat{OR}$ se debe al hecho que el rango de sus posibles valores va de 0 a ∞, con el valor nulo igual a 1. En teoría, para tamaños de muestra suficientemente grandes, la distribución de $\hat{OR}$ es normal. Desafortunadamente, este requerimiento de tamaño de muestra excede al de la mayoría de estudios. Por lo tanto las inferencias están usualmente basadas

en la distribución muestral para $Ln\left(\hat{OR}\right) = \hat{\beta}_1$, lo cual tiende a seguir una distribución normal para tamaños de muestra mucho más pequeños.

Un intervalo confidencial del $100(1-\alpha)\%$ para el OR se obtiene calculando los puntos extremos de un intervalo confidencial para el coeficiente β_1, y luego exponenciamos estos valores. En general, los puntos extremos están dados por la expresión

$$\exp\left[\hat{\beta}_1 \pm \hat{Z}_{1-\alpha/2} \times \hat{SE}\left(\hat{\beta}_1\right)\right]$$

VARIABLE INDEPENDIENTE POLITÓMICA

Supongamos que la variable independiente tiene $K > 2$ valores distintos, la escala de medición es nominal. Por lo tanto, debemos formar un conjunto de variables diseño para representar las categorías de la variable. El método para especificar las variables diseño implica igualar a 0 las categorías del grupo de referencia, y luego fijar una variable diseño simple igual a 1 para cada uno de los siguientes grupos. Por ejemplo, la variable RAZA generará las variables diseño siguientes

RAZA (Categorías)	Variables Diseño		
	RAZA 1	RAZA 2	RAZA 3
Blanco	0	0	0
Negro	1	0	0
Mestizo	0	1	0
Otro	0	0	1

Este método de codificación (indicador) es el método por defecto en la mayoría de softwares estadísticos. Una vez construidas las variables diseño el procedimiento a seguir para la interpretación es el mismo que si se tratase de una variable independiente dicotómica.

VARIABLE INDEPENDIENTE CONTÍNUA

En este caso la interpretación del coeficiente estimado depende de cómo ha sido ingresado en el modelo y las unidades particulares de la variable. Para propósitos de desarrollar el método para interpretar el coeficiente para una variable continua asumimos que el logit es lineal en la covariable continua X, la ecuación para el logit es $g(X) = \beta_0 + \beta_1 X$, el coeficiente pendiente, β_1, da el cambio en el Log de odds para un incremento de "1" unidad en X, esto es, $\beta_1 = g(X+1) - g(X)$ para cualquier valor de X. Mayormente el valor de "1" no es clínicamente interesante. Por ejemplo, 1 año de incremento en la edad puede ser demasiado pequeño para ser considerado importante, un cambio de 10 años podría ser considerado más útil. De otro lado, si el rango de X va de 0 a 1, entonces un cambio de 1 es demasiado grande y un cambio de 0.01 puede ser más realista.

Por lo tanto, para proveer una interpretación útil de la covariable de escala continua, necesitamos desarrollar un método para estimación puntual y por intervalos para un cambio arbitrario de "c" unidades en la covariable.

El Log de odds ratio para un cambio de "c" unidades en X se obtiene de la diferencia logit $g(X+c) - g(X) = c\beta_1$ y el odds ratio asociado se obtiene exponenciando esta diferencia logit, $OR(C) = OR(X+c, X) = \exp(c\beta_1)$. Una

estimación se puede obtener remplazando β_1 con su estimador máximo verosímil $\hat{\beta}_1$.

Una estimación del error estándar necesario para la estimación del intervalo de confianza se obtiene multiplicando el error estándar de β_1 por c. Por lo tanto, los límites del IC $100(1-\alpha)\%$ de $OR(c)$ son:

$$\exp\left[c\,\hat{\beta}_1 \pm Z_{1-\alpha/2}\, c \times \hat{SE}\left(\hat{\beta}_1 \right) \right]$$

Debido a que tanto la estimación puntual como la estimación por intervalos dependen de la elección de c, este debería estar claramente especificado en todas las tablas y cálculos que se presenten. La elección de c puede ser difícil para algunos, por ejemplo ¿Por qué usar un cambio de 10 años cuando 5 o 15 años pueden ser igualmente convenientes?. Podríamos usar algún valor razonable, pero el propósito que debe tenerse en cuenta es proveer al lector de nuestro análisis de una clara indicación de cómo el riesgo de la variable respuesta presente cambia con la variable en cuestión. En general, los cambios en múltiplos de 5 ó 10 pueden ser significativos y fácilmente comprendidos.

1.4 EVALUACIÓN DEL MODELO

Siempre que se construye un modelo de regresión es fundamental, antes de pasar a extraer conclusiones, el corroborar que el modelo calculado se ajusta efectivamente a los datos usados para estimarlo.

1.4.1 PRUEBAS PARA LA SIGNIFICANCIA DE LOS COEFICIENTES DEL MODELO

Después de estimar los coeficientes del modelo, debemos evaluar la significancia de las variables en el modelo. Esto involucra la formulación y prueba de una hipótesis estadística para determinar si las variables independientes en el modelo están significativamente relacionadas a la variable respuesta. El método para desarrollar esta prueba es bastante general y difiere de un tipo de modelo a otro, sólo en detalles específicos.

1.4.1.1 TEST DE RAZÓN DE VEROSIMILITUD

Los test para la significancia de una variable en cualquier modelo se relacionan con la pregunta ¿El modelo que incluye la variable a evaluar nos dice más acerca de la variable respuesta que un modelo que no incluye esa variable? Respondemos esa pregunta comparando los valores observados de la variable respuesta con los valores predichos para cada uno de los dos modelos, el primero con y el segundo sin la variable a evaluar. Es importante mencionar que no estamos evaluando si los valores predichos son una representación exacta de los valores observados en sentido absoluto (sería Bondad de ajuste), sino más bien nuestra pregunta es planteada en sentido relativo.

En regresión lineal, la significancia del coeficiente pendiente se evalúa con una tabla de análisis de varianza, la cual particiona la suma total de desviaciones al cuadrado de las observaciones respecto a su media en dos partes:

(1)La suma de desviaciones al cuadrado de las observaciones respecto a la línea de regresión SSE (suma de cuadrados residual).

(2)La suma de cuadrados de valores predichos basados en el modelo de regresión, respecto a la media de la variable dependiente SSR (suma de cuadrados de la regresión).

Esto es una forma conveniente de comparar los valores observados y los valores predichos bajo los dos modelos, esta comparación se basa en el cuadrado de la distancia entre los dos. Si Y_i es el valor observado y $\hat{Y}_i$ el valor predicho para el i-ésimo individuo bajo el modelo, entonces la estadística usada para evaluar esta comparación es:

$$SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}i)^2$$

Bajo el modelo que no contiene la variable independiente a evaluar el único parámetro es $\beta_0=\overline{Y}$, la media de la variable respuesta.

En este caso $\hat{Y}i=\overline{Y}$ y SSE es igual a la varianza total. Cuando incluimos la variable independiente en el modelo y disminuye SSE se deberá al hecho que el coeficiente pendiente para la variable no es 0. El cambio en el valor se SSE se debe a la fuente de variabilidad de la regresión (SSR).

$$SSR=\left[\sum_{I=1}^{n}\left(Y_I-\bar{Y}_i\right)^2\right]-\left[\sum_{I=1}^{n}\left(Y_I-\hat{Y}_i\right)^2\right]$$

En regresión lineal el interés se centra en el tamaño de SSR, un valor grande indica que la variable independiente es importante, mientras que lo contrario sugiere que la variable independiente no es útil en predecir la respuesta.

Con regresión logística se sigue el mismo principio, comparar los valores observados de la variable respuesta con los valores predichos obtenidos de los modelos con y sin la variable en cuestión, esta comparación se basa en la función logaritmo de la verosimilitud

$$L(\beta) = \ln[\lambda(\beta)] = \sum_{i=1}^{n} \{Y_i \ln[\Pi(X_i)] + (1 - Y_i) \ln[1 - \Pi(X_i)]\}$$

Para entender esta comparación, debemos pensar en un valor observado de la variable respuesta como un valor predicho resultante de un modelo saturado; un modelo saturado es aquel que contiene tantos parámetros como datos hay.

La comparación de valores observados y predichos usando la función verosimilitud se basa en la siguiente expresión:

$$D = -2 \ln \left[\frac{(Verosimilitud..del..\text{mod}elo..ajustado)}{(Verosimilitud..del..\text{mod}elo..saturado)} \right]$$

La cantidad dentro de los corchetes en la expresión de arriba se llama razón de verosimilitud. Usar -2 veces su logaritmo es necesario para obtener una cantidad

cuya distribución es conocida y puede ser usada para propósitos de probar la hipótesis:

H0: $\beta_0 = \beta_1 = \ldots = \beta_p = 0$

H1: $\beta_j \neq 0$ para algún j = 1,2,…,p

Tal test se llama el “Test Razón de Verosimilitud”, remplazando las verosimilitudes respectivas en la expresión anterior tenemos:

$$D = -2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\hat{\Pi}_i}{Y_i}\right) + (1-Y_i)\ln\left(\frac{1-\hat{\Pi}_i}{1-Y_i}\right)\right]$$

donde $\Pi_i = \prod(X_i)$.

La estadística se llama la desvianza, y cumple el mismo rol que la suma de cuadrados de residuos en regresión lineal.

La desvianza de un modelo de investigación compara el logaritmo de verosimilitud del modelo de interés con el logaritmo de la verosimilitud del modelo saturado, es decir, aquel que se ajusta completamente a los datos (para cada observación existe un parámetro).

Desvianza pequeña: La explicación del modelo ajustado es prácticamente igual a la del modelo saturado, es decir, podemos usar el modelo ajustado pues generalmente tiene menos parámetros.

Desvianza grande: La explicación del modelo ajustado es pobre, es decir, no podemos usar el modelo ajustado.

Por definición de modelo saturado $\hat{\Pi}_i = Y_i$, y

$$L(\text{mod}\, elo..saturado) = \prod_{i=1}^{n} Y_i^{Y_i} (1 - Y_i)^{(1-Y_i)} = 1$$

Por lo tanto

D = - 2 ln (Verosimilitud del modelo ajustado)

Para evaluar la significación de una variable independiente comparamos el valor de D con y sin la variable independiente en la ecuación. El cambio en D debido a la inclusión de la variable independiente en el modelo se obtiene como

G = D(modelo sin la variable) – D(modelo con la variable)

Esta estadística juega el mismo rol en regresión logística que el numerador del test F parcial en regresión lineal. Debido a que la verosimilitud del modelo saturado es común para ambos valores de D, G puede expresarse como:

$$G = -2 \ln \left[\frac{Verosimilitud.\text{sin}.la.\text{var}iable}{Verosimilitud.con..la.\text{var}iable} \right]$$

Para el caso específico de una variable independiente simple, es fácil mostrar que cuando la variable no está en el modelo, la estimación máximo verosímil de β_0 es $\text{Ln}\left(\frac{n_1}{n_0}\right)$, donde: $n_1 = \sum Y_i$ y $n_1 = \sum (1 - Y_i)$ y el valor predicho es constante, $\frac{n_1}{n}$. En este caso, el valor de G es:

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\Pi}_i^{Y_i} \left(1 - \hat{\Pi}_i\right)^{1-Y_i}} \right]$$

ó $$G = 2\left\{ \sum_{i=1}^{n} \left[Y_i \ln\left(\hat{\Pi}_i\right) + (1 - Y_i) \ln\left(1 - \hat{\Pi}_i\right) \right] - \left[n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n) \right] \right\}$$

Bajo la hipótesis $\beta_1 = 0$, la estadística G sigue una distribución Chi-cuadrado con 1 grado de liberad.

Para el caso multivariable, los valores ajustados, $\hat{\Pi}$, bajo el modelo están basados en el vector que contiene p+1 parámetros $\hat{\beta}$. Bajo la hipótesis que los p coeficientes "pendiente" para las covariables en el modelo son iguales a 0, la distribución de G será Chi-cuadrado con p grados de libertad.

1.4.1.2 TEST DE WALD

Para el caso univariable, este test se obtiene comparando la estimación máximo verosímil del parámetro pendiente, $\hat{\beta}_1$, con la estimación de su error estándar.

La razón resultante, bajo la hipótesis que β_1=0, seguirá una distribución normal estándar.

$$W = \frac{\hat{\beta}_1}{\hat{SE}\left(\hat{\beta}_1\right)}$$

Hauck y Donner (1977) examinaron el comportamiento del test de wald y encontraron que frecuentemente falla en rechazar la hipótesis nula cuando el coeficiente era significativo. Ellos recomiendan usar el test de razón de verosimilitud.

Para el caso multivariable el test de wald se obtiene de la siguiente manera:

$$W = \hat{\beta}'\left[\hat{Var}\left(\hat{\beta}\right)\right]^{-1}\hat{\beta} = \hat{\beta}'\left(XVX\right)\hat{\beta}$$

El cual se distribuye como una Chi-cuadrado con p+1 grados de libertad bajo la hipótesis que los p+1 coeficientes son igual a 0.

Los test solo para los p coeficientes pendiente se obtienen eliminando $\hat{\beta}_0$ de $\hat{\beta}'$ y la fila correspondiente (1ra o última) y columna (1ra o última) de (X'VX).

Debido a que la evaluación de este test requiere la capacidad para desarrollar cálculos matriciales y obtener $\hat{\beta}$, su uso no es preferido al test de razón de verosimilitud de la significación del modelo.

1.4.1.3 TEST SCORE

Para el caso univariable este test se basa en la distribución condicional de la derivada en la ecuación (1.2.1):

$$\sum X_{ij}\left[Y_i - \Pi(X_i)\right] = 0$$

Dada la derivada en la ecuación (1.2.2)

$$\sum \left[Y_i - \Pi(X_i)\right] = 0$$

El test usa el valor de la ecuación (1.2.2) calculada usando β_0 = Ln$\left(\frac{n_1}{n_0}\right)$ y β_1 =0, bajo estos valores de los parámetros $\hat{\Pi}$ = $\frac{n_1}{n} = \overline{Y}$. Así el lado izquierdo de la ecuación (1.2.2) es $\sum X_i$ (Y_i - $\overline{Y}$).

Puede demostrarse que la varianza estimada es $\overline{Y}$ (1-$\overline{Y}$) $\sum\left(X_i - \overline{X}\right)^2$. El estadístico de prueba para el test score es:

$$ST = \frac{\sum_{i=1}^{n} X_i\left(Y_i - \overline{Y}\right)}{\sqrt{\overline{Y}\left(1-\overline{Y}\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

Para el caso multivariable, este test se basa en la distribución teórica de las p derivadas del logaritmo de verosimilitud L(β) con respecto a β. El cálculo de este test tiene la misma dificultad que el test de Wald.

1.4.2 BONDAD DE AJUSTE DEL MODELO

Una vez que hemos construido el modelo de regresión logística, el siguiente paso es conocer cuan efectivamente el modelo describe la variable respuesta, es decir, la bondad de ajuste.

Sea $Y^{I} = (Y_1, Y_2, ..., Y_n)$ valores muestrales observados de la variable respuesta en forma vectorial y $\hat{y}' = \left(\hat{Y}_I, \hat{Y}_2, ..., \hat{Y}_n \right)$ valores predichos por el modelo o valores ajustados.

Decimos que ha y un buen ajuste del modelo, sí: (1) Las medida resumen de la distancia entre Y y $\hat{Y}$ son pequeñas, y (2) La contribución de cada par $(Y_i, \hat{Y}_I)$, i = 1,2,...,n a las medidas resumen no es sistemática y es pequeña en relación a la estructura de error del modelo.

Por lo tanto, un análisis completo del ajuste del modelo involucra el calculo de medidas resumen de la distancia entre Y y $\hat{Y}$, y el examen a través de los componentes individuales de estas medidas.

Cuando la etapa de ajuste del modelo ha terminado, una serie de pasos pueden usarse para asegurar el buen ajuste del modelo, entre ellos tenemos:

1. Cálculo y evaluación de medidas de ajuste globales.

2. Examen de los componentes individuales de las estadísticas de resumen, con frecuencia gráficamente.

3. Examen de otras medidas de la diferencia o distancia entre los componentes de Y y $\hat{Y}$.

MEDIDAS RESUMEN DE BONDAD DE AJUSTE

Las estadísticas resumen, por naturaleza, no pueden darnos información acerca de loa componentes individuales del modelo. Una valor pequeño para una de estas estadísticas, no descarta la posibilidad de alguna desviación de ajuste sustancial para pocos sujetos. De otro lado un valor grande para una de estas estadísticas es una clara evidencia de un problema sustancial con el modelo.

1.4.2.1 ESTADÍSTICA CHI-CUADRADO DE PEARSON Y DESVIANZA

En regresión lineal, las medidas resumen de ajuste son funciones de un residual definido como la diferencia entre los valores observados y ajustados. En regresión logística hay diferentes formas para medir la diferencia entre los valores observados y ajustados, los valores ajustados son calculados para cada patrón covariable y dependen de la probabilidad estimada para ese patrón covariable, denotemos el valor ajustado para el j-ésimo patrón covariable como $\hat{Y}_j$, donde:

$$\hat{Y}_j = m_j \hat{\Pi}_j = m_j \left[\frac{e^{\hat{g}(X_j)}}{1 + e^{\hat{g}(X_j)}} \right]$$

Donde $\hat{g}(X_j)$ es el logit estimado.

Para un patrón covariable particular, el residual de Pearson se define como:

$$r\left(Y_j,\hat{\Pi}_j\right)=\frac{\left(Y_j-m_j\hat{\Pi}_j\right)}{\sqrt{m_j\hat{\Pi}_j\left(1-\hat{\Pi}_j\right)}} \qquad (1.4.2.1.1)$$

La estadística resumen basada en estos residuales es la estadística chi- cuadrado de Pearson

$$\chi^2=\sum_{j=1}^{J}r\left(Y_j,\Pi_j\right)^2$$

La Desvianza residual, se define como:

$$d\left(Y_j,\hat{\Pi}_j\right)=\pm\left\{2\left[Y_j\ln\left(\frac{Y_j}{m_j\hat{\Pi}_j}\right)+\left(m_j-Y_j\right)\ln\left[\frac{\left(m_j-Y_j\right)}{m_j\left(1-\hat{\Pi}_j\right)}\right]\right]\right\}^{1/2} \qquad (1.4.2.1.2)$$

donde el signo + ó - , es el mismo signo de $\left(Y_j-m_j\hat{\Pi}_j\right)$. Para patrones covariable con $Y_j=0$ la desvianza residual es:

$$d\left(Y_j,\hat{\Pi}_j\right)=-\sqrt{2m_j\left|\ln\left(1-\hat{\Pi}_j\right)\right|}$$

y la desvianza residual cuando $Y_j=m_j$, es:

$$d\left(Y_j,\hat{\Pi}_j\right)=\sqrt{2m_j\left|\ln\left(\hat{\Pi}_j\right)\right|}$$

La estadística resumen basada en las desvianzas residuales es la Desvianza

$$D = \sum_{j=1}^{J} d\left(Y_j, \hat{\Pi}_j \right)^2$$

La distribución de las estadísticas χ^2 y D bajo la suposición que el modelo ajustado es correcto en todos los aspectos es Chi-cuadrado con grados de libertad igual a J-(p+1). Para la desvianza esta afirmación se deduce del hecho que D es la estadística de prueba razón de verosimilitud de un modelo saturado con J parámetros versus el modelo ajustado con p+1 parámetros. Similar teoría provee la distribución nula de χ^2. El problema es que cuando $J \approx n$, la distribución se obtiene bajo n-asintótico, y por lo tanto el número de parámetros se incrementa en la misma razón que el tamaño muestral. Así , los p-valores calculados para estas dos estadísticas cuando $J \approx n$, usando la distribución $\chi^2_{(J-p-1)}$, son incorrectos.

Una forma de evitar esta dificultad es agrupar los datos de tal forma que m-asintótico pueda ser usado. Para entender el razonamiento detrás de las varias estrategias de agrupamiento que han sido propuestos, es útil pensar en la estadística χ^2 como la estadística Pearson y en la estadística D como la estadística Chi-cuadrado Razón de Verosimilitud que resultan de una tabla 2XJ. Las filas de la tabla correspondientes a los dos valores de la variable respuesta, Y=1,0. Las J columnas corresponden a los J posibles patrones covariable. La estimación del valor esperado bajo la hipótesis que el modelo logístico en cuestión es el modelo correcto para las celdas correspondientes a la fila Y=1 y a la j-ésima columna es $m_j(1-\hat{\Pi}_j)$. Las estadísticas χ^2 y D se calculan de esta tabla en la forma usual.

Pensar en las estadísticas como resultado de una tabla de 2XJ da alguna idea intuitiva de porque no podemos esperar que ellas sigan la distribución $\chi^2_{(J-p-1)}$. Cuando los test chi-cuadrado son calculados de una tabla de contingencia, los p-valores son correctos bajo la hipótesis nula si los valores esperados estimados son "grandes" en cada celda. Esta condición se mantiene bajo m-asintótico. Aunque esta es una sobre-simplificación del problema, es esencialmente correcto. En la tabla 2XJ descrito arriba los valores esperados son siempre bastante pequeños debido a que el número de columnas se incrementa cuando n se incrementa. Para evitar este problema podemos reducir las columnas a un número fijo de grupos, g, y luego calcular las frecuencias esperadas y observadas. Fijando el número de columnas, las frecuencias esperadas estimadas se hacen grandes a medida que n se hace grande. Así, m-asintótico se mantiene.

1.4.2.2 ESTADÍSTICA DE HOSMER Y LAMESHOW

En el caso de la regresión logística una idea bastante intuitiva es calcular la probabilidad de aparición del suceso, para todos los integrantes de la muestra. Si el ajuste es bueno, es de esperar que un valor alto de probabilidad se asocie con presencia real del suceso, y viceversa, si el valor de esa probabilidad calculada es bajo, cabe esperar también ausencia del suceso.

Este procedimiento consiste básicamente en dividir el recorrido de la probabilidad en deciles de riesgo (esto es por ejemplo, probabilidad de hipertensión < 0.1, < 0.2, y así hasta <1) y calcular tanto la distribución de hipertensos, como no hipertensos

prevista por la ecuación y los valores realmente observados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de chi² .

Esta estadística compara el número observado con el número esperado de éxitos de g grupos formados.

El primer grupo deberá contener n_i elementos correspondientes a las n_1 menores probabilidades ajustadas, las cuales se denotan como $\hat{\Pi}(1) \leq \hat{\Pi}(2) \leq ..... \leq \hat{\Pi}\left(n_1'\right)$. El segundo grupo deberá contiene los n_2 elementos correspondientes a las siguientes probabilidades ajustadas, $\hat{\Pi}\left(n_1'+1\right) \leq \hat{\Pi}\left(n_1'+2\right) \leq ..... \leq \hat{\Pi}\left(n_1'+n_2'\right)$. Y así sucesivamente, hasta el último grupo que deberá contener las n_g' mayores probabilidades ajustadas $\hat{\Pi}\left(n_1'+...+n_{g-1}'+1\right) \leq \hat{\Pi}\left(n_1'+...+n_{g-1}'+2\right) \leq ..... \leq \hat{\Pi}(n)$.

El número de éxitos del primer grupo formado será $O_1 = \sum_{j=1}^{n_1'} Y_{(j)}$, en el cual $Y_{(j)}$ = 0 para un elemento correspondiente al fracaso y $Y_{(j)}$ = 1 si es éxito.

Generalizando, tenemos: $O_i = \sum_{j=n_1'+...+n_{i-1}'+1}^{n_1'+...+n_i'} Y_{(j)}$, $2 \leq i \leq g$. La estadística se define como:

$$\hat{C} = \sum_{i=1}^{g} \frac{\left(O_i - n_1' \hat{\Pi}_i\right)^2}{n_1' \Pi_i \left(1 - \hat{\Pi}_i\right)}$$

Donde $\hat{\Pi}_1 = \left(1/n_1'\right) \sum_{j=1}^{n_1'} \hat{\Pi}_{(j)}$ y $\hat{\Pi}_i = \left(1/n_i'\right) \sum_{j=n_1'+...+n_{i-1}'+1}^{n_1'+...+n_i'} \hat{\Pi}_{(j)}$, $2 \leq i \leq g$. Hosmer-Lemeshow sugieren la formación de g =10 grupos del mismo tamaño (aproximadamente), de modo que el primer grupo contenga n_i elementos

correspondientes a las (n/10) menores probabilidades ajustadas y así sucesivamente hasta ajustar el último grupo con n_{10} elementos correspondientes a las (n/10) mayores probabilidades ajustadas.

Cuando no hay empates, esto es, n_i =1, para todo i, es relativamente fácil construir los 10 grupos con tamaños aproximadamente iguales, más no sucede igual cuando hay empates, puede ser necesario que dos individuos con la misma configuración de covariables sean colocados en grupos adyacentes, a fin de que los grupos formados no tengan tamaños muy desiguales.

Hosmer-Lemeshow demostraron que la distribución nula asintótica de C puede aproximarse a una distribución Chi-cuadrado con (g-2) grados de libertad.

1.4.3 DIAGNÓSTICO DEL MODELO

Las estadísticas resumen basadas en los residuales Chi-cuadrado de Pearson proveen un simple número que resume la concordancia entre valores observados y ajustados, la ventaja (así como desventaja) de esta estadística es que un simple número resume considerable información. Por lo tanto, antes de concluir que el modelo "ajusta", es crucial que otras medidas sean examinadas para ver si el ajuste se mantiene sobre todo el conjunto de patrones covariable. Esto se efectúa a través de una serie de medidas especializadas que caen bajo la denominación general de diagnósticos de regresión.

Asumamos que el modelo ajustado contiene p covariables y que ellos forman J patrones covariable. En regresión lineal la suposición clave es que la varianza del error no depende de la media condicional, $E(Y_j / X_j)$. Sin embargo, en regresión logística tenemos errores binomiales y la varianza del error es una función de la media condicional

$$Var(Y_j / X_j) = m_j E(Y_j / X_j) \times [1 - E(Y_j / X_j)] = m_j \pi(X_j)[1 - \pi(X_j)]$$

Así, iniciamos con residuales como el de Pearson y la desvianza, los cuales han sido divididos por estimaciones de su error estándar , esto puede no ser completamente obvio en el caso de la desvianza residual. Sea r_j y d_j los valores de las expresiones dadas en las ecuaciones (1.4.2.1.1) y (1.4.2.1.2) respectivamente, para patrones covariable X_j. Debido a que cada residual ha sido dividido por una estimación aproximada de su error estándar, esperamos que si el modelo de

regresión logístico es correcto, estas cantidades tengan aproximadamente una media igual a 0 y una varianza igual a 1.

Además de los residuales para cada patrón covariable, otras medidas importantes para la formación de diagnósticos de regresión lineal son la matriz sombrero y los valores "Leverage" derivados de este. En regresión lineal, la matriz sombrero es H = X(X'X)$^{-1}$X'; por ejemplo, $\hat{Y} = HY$. Los residuales de regresión lineal, $\left(Y - \hat{Y}\right)$, expresados en términos de la matriz sombrero son $(I - H)Y$, donde I es la matriz identidad $J \times J$.

Usando regresión lineal de mínimos cuadrados ponderados como un modelo, Pregibon (1981) derivó una aproximación lineal para los valores ajustados, los cuales producen una matriz sombrero para la regresión logística. Esta matriz es:

$$H = V^{1/2} X\left(X'VX\right)^{-1} X'V^{1/2} \qquad (1.4.3.1)$$

donde V es una matriz diagonal $J \times J$ con elementos generales

$$V_j = m_j \hat{\Pi}\left(X_j\right)\left[1 - \hat{\Pi}\left(X_j\right)\right]$$

En regresión lineal los elementos de la diagonal de la matriz son llamados los valores Leverage y son proporcionales a la distancia de X_j a la media de los datos. Este concepto de distancia a la media es importante en regresión lineal; los puntos que están lejos de la media pueden tener considerable influencia en los valores de los parámetros estimados, la extensión del concepto de Leverage para regresión logística requiere discusión adicional.

Sea la cantidad h_j, que denota el j-ésimo elemento diagonal de la matriz H definida en la ecuación (1.4.3.1), se puede demostrar que

$$h_j = m_j \hat{\Pi}(X_j)\left[1 - \hat{\Pi}(X_j)\right] X_j' \left(X'VX\right)^{-1} X_j' = v_j \times b_j$$

donde $b_j = X_j' \left(X'VX\right)^{-1} X_j'$ y $X_j' = \left(1, X_{1j}, X_{2j}, ..., X_{pj}\right)$ es el vector de valores covariable definidos en el j-ésimo patrón covariable. La suma de los elementos de la diagonal de H es, como en el caso de regresión lineal, $\sum h_j = (p+1)$, el número de parámetros en el modelo. En regresión lineal la dimensión de la matriz es usualmente $n \times n$ y así ignora cualquier patrón covariable común en los datos. Con esta formulación, cualquier elemento diagonal en la matriz sombrero tiene un limite superior de 1/k, donde k es el número de sujetos con el mismo patrón covariable. Si formulamos la matriz sombrero para regresión logística como una matriz $n \times n$, entonces cada elemento diagonal esta limitado superiormente por $1/m_j$, donde m_j es el número total de sujetos con el mismo patrón covariable. Cuando la matriz sombrero se basa en datos agrupados por patrones covariable, el limite superior para cualquier elemento de la diagonal es 1.

Otra estadística de diagnóstico útil es la que examina el efecto que tiene eliminar todos los sujetos con un patrón covariable particular en el valor de los coeficientes estimados y en la medida de ajuste global, χ^2 y D. El cambio en el valor de los coeficientes estimados es análogo a la medida propuesta por Cook (1977,1979) para regresión lineal. Este se obtiene como la diferencia estandarizada entre $\hat{\beta}$ y $\hat{\beta}_{(-j)}$, donde estos representan las estimaciones máximo verosímiles calculadas

usando todos los J patrones covariable y excluyendo a los m_j sujetos con patrón X_j respectivamente, y estandarizando vía la matriz covarianza estimada de $\hat{\beta}$. Pregibon (1981) demostró, para una aproximación lineal, que esta cantidad para regresión logística es

$$\Delta\hat{\beta}_j = \left(\hat{\beta} - \hat{\beta}_{(-j)}\right)' \left(X'VX\right) \left(\hat{\beta} - \hat{\beta}_{(-j)}\right) = \frac{r_j^2 h_j}{(1-h_j)^2} = \frac{r_{sj}^2 h_j}{(1-h_j)}$$

Usando similares aproximaciones lineales puede demostrarse que la disminución en el valor de la estadística Chi-cuadrado de Pearson debida a la eliminación de los sujetos con patrón covariable X_j es

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2 \qquad (1.4.3.2)$$

Una similar cantidad puede obtenerse para el cambio en la desvianza

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}$$

Sí remplazamos r_j^2 por d_j^2, este produce la aproximación

$$\Delta D_j = \frac{d_j^2}{(1-h_j)} \qquad (1.4.3.3)$$

la cual tiene forma similar a la expresión en la ecuación (1.4.3.2). Estas estadísticas diagnóstico son bastante atractivas conceptualmente, ellas permiten identificar aquellos patrones covariable que tienen un ajuste pobre (valores grandes de ΔX_j^2 y/o ΔD_j), y aquellos que tienen mucha influencia en los valores de los parámetros

estimados (valores grandes de $\Delta\hat{\beta}_j$. Después de identificar estos patrones de influencia, podemos empezar a determinar el rol que juegan en el análisis. Para interpretar el valor de las estadísticas de diagnóstico usamos el enfoque gráfico, los valores grandes de diagnóstico o aparecen como puntos o se ubican en las esquinas extremas del gráfico. Un valor de la estadística diagnóstico que cae fuera del nivel de los puntos se dice que es extremo si excede algún percentil de su distribución.

En la práctica, la determinación de "grande" es un juicio basado en la experiencia y el conjunto particular de datos que están siendo analizados. Usar la $N(0,1)$, o equivalentemente, la distribución $\chi^2(1)$ para cantidades al cuadrado puede proveer una guía de lo que es grande, sin embargo, estos percentiles deben usarse con extrema precaución. No hay sustituto para la experiencia en el uso efectivo de las estadísticas diagnóstico.

La clave del análisis de diagnóstico consta de los siguientes gráficos:

(1) ΔX_j^2 vs. $\hat{\Pi}_j$

(2) ΔD_j vs. $\hat{\Pi}_j$

(3) $\Delta\hat{\beta}_j$ vs. $\hat{\Pi}_j$

Otros gráficos algunas veces útiles son:

(4) ΔX_j^2 vs. h_j

(5) ΔD_j vs. h_j

(6) $\Delta \hat{\beta}_j$ vs. h_j

1.5 TABLAS DE CLASIFICACIÓN

Una forma intuitiva de resumir los resultados de un modelo de regresión logístico ajustado es a través de una tabla de clasificación. Esta tabla es el resultado de la clasificación cruzada de la variable respuesta, Y, con una variable dicotómica cuyos valores se derivan de las probabilidades logísticas estimadas.

Para obtener la variable dicotómica derivada debemos definir un punto de corte, C, y comparar cada probabilidad estimada con C. Si la probabilidad estimada excede a C, entonces hacemos que la variable derivada sea igual a 1, en caso contrario esta es igual a 0. el valor más comúnmente usado para C es 0.5. Apelar a este tipo de enfoque para la valoración del modelo viene de la relación cercana de la regresión logística al análisis discriminante cuando la distribución de las covariables es normal multivariable dentro de los dos grupos resultantes.

En este enfoque, las probabilidades estimadas son usadas para predecir a los miembros del grupo.

SENSIBILIDAD, es la probabilidad de obtener verdaderos positivos a partir de un criterio ó regla de decisión establecida. Se estima mediante la proporción de casos de riesgo, que son clasificados como tales.

ESPECIFICIDAD, es la probabilidad de obtener verdaderos negativos a partir de un criterio ó regla de decisión establecida. Se estima mediante la proporción de casos control ó sin riesgo, que son clasificados como tales.

En resumen, la especificidad se refiere a la capacidad del modelo desarrollado, y de su regla de decisión asociada, de distinguir los casos que no están en riesgo. La sensibilidad en cambio representa la capacidad del procedimiento para detectar los casos que se encuentran en situación de riesgo. Ambas medidas están relacionadas y dependen del punto de corte o umbral en el indicador cuantitativo, de tal manera que si éste es muy bajo, tendremos alta especificidad pero baja sensibilidad, y si, por el contrario establecemos un umbral alto tendremos lo contrario.

Presumiblemente, si el modelo predice con exactitud a los miembros del grupo de acuerdo a algún criterio, entonces se piensa que provee evidencia de que el modelo ajusta, desafortunadamente, este no siempre puede ser el caso. Por ejemplo, es fácil construir una situación donde el modelo de regresión logístico es el modelo correcto y ajusta, pero la clasificación es pobre.

La exactitud o inexactitud de la clasificación no direcciona nuestro criterio para bondad de ajuste: que la distancia entre valores observados y esperados sea no sistemática y dentro de la variación del modelo. Sin embargo, la tabla de clasificación puede ser útil junto a otras medidas basadas más directamente en residuales.

Una importante razón de porque las medidas derivadas de una tabla de clasificación 2X2 (sensibilidad y especificación) no deberían usarse como medida de bondad de ajuste del modelo es que ellos dependen pesadamente de la distribución de probabilidad de la muestra.

La tabla de clasificación es más apropiada cuando la clasificación es un propósito establecido del análisis; en caso contrario este debería solo completar a métodos más rigurosos de bondad de ajuste.

CAPÍTULO II

FACTORES ASOCIADOS A LA PRESENCIA DE ASMA EN ESCOLARES DE LA CIUDAD DE MOQUEGUA

2.1 PLANTEAMIENTO DEL PROBLEMA

El asma constituye uno de los problemas de salud que afecta a la mayoría de niños en edad escolar. Con el propósito de cuantificar el problema causado por esta enfermedad, el Ministerio de Salud realizó el estudio "Encuesta Sobre Enfermedades Respiratorias - 2004", llevada a cabo en la ciudad de Moquegua; los resultados del presente trabajo serán importantes por que se traducirán en recomendaciones útiles para las políticas a seguir por dicha institución.

En el Perú se han incrementado las investigaciones sobre epidemiología del asma desde hace 5 años[1]. Hoy se sabe que uno de cada cinco consultantes a los servicios de salud tiene antecedentes de asma o sinónimos y uno de cada 10, por lo menos, tiene asma actualmente.

En Perú, se estima que la prevalencia de asma (todos los casos de asma que existen en un período) es de 10% o más. Desde 1998 los niños en edad pre-escolar y escolar tienen acceso a medicación antiasmática en Perú, a través de los Programas de salud implementados en los establecimientos del Ministerio de Salud (inhaladores broncodilatadores y antiinflamatorios)[1]. Esto es un avance social significativo.

La problemática presentada, motivó la necesidad de investigar los factores asociados a la presencia de asma; para llevar a cabo este estudio se consideró la ciudad de Moquegua.

El asma es una enfermedad inflamatoria crónica de las vías respiratorias en la que interfieren múltiples células, en particular los mastocitos, eosinófilos, linfocitos T, neutrófilos y células epiteliales. En individuos susceptibles esta inflamación causa

[1] El Boletín Electrónico **LIBERTAD *para respirar,*** informativo de la Unidad de Control de Asma, Ambiente y Tabaco (UCAAT), del Hospital Nacional Dos de Mayo, Agosto 2002 Lima, Perú.

episodios de sibilancias, disnea, opresión torácica y tos, especialmente por la noche y/o primeras horas de la mañana. Estos síntomas suelen asociarse a una limitación variable del flujo aéreo, reversible de forma espontánea o con tratamiento.

La definición de asma en niños menores de tres años es particularmente difícil porque las sibilancias en la mayoría de los niños de este grupo son debidas, principalmente, a factores mecánicos asociados con el tamaño de la vía respiratoria. Durante los últimos 10 años ha habido un gran debate conceptual o terminológico para describir las sibilancias episódicas que ocurren al principio de la niñez y que se desencadenan en muchas ocasiones, por una infección viral. Se han manejado términos como bronquitis espástica, catarros descendentes, bronquitis asmatiforme, bronquitis sibilante, bronquitis disneizante, asma infecciosa, sibilancias postbronquiolitis o síndrome postbronquiolitis, lo que implica una reacción causal del episodio, aunque no siempre puede ser demostrado. Más tarde estos términos fueron sustituidos por otros como hiperreactividad bronquial, que no es un diagnóstico sino una característica o situación del bronquio. Otros autores anglosajones han propuesto el término descriptivo de Lower Respitatory Illness (LRI), enfermedad de vías respiratorias inferiores con o sin sibilancias. Al final, intentando globalizar el concepto se decidió llamarlo simplemente asma.

El asma y, en sentido amplio las enfermedades alérgicas, presentan una agregación familiar[2] que sugiere la existencia de una base genética. La relación entre el asma de los padres y de los hijos es bastante conocida y, a pesar de que

[2] Ballesta F. Genética y alergia. Allergol Immunopathol 1998; 26: 83-86. Postma DS, Bleecker ER, Amelung PJ, Holroyd KJ, Xu J, Panhuysen C et al. Genetic susceptivility to asthma– bronquial hyperresponsiveness coinherited with a mayor gene for atopy. N Engl J Med 1995; 5: 894-900.

hay estudios en los que se pone de manifiesto que esta relación es independiente del sexo de los padres, parece que el asma materna desempeña el papel más importante. A pesar de la importancia de los factores genéticos, en la etiopatogenia del asma se propone un mecanismo poligénico multifactorial, en el que se acepta la suma de factores genéticos y ambientales.

Se ha comprobado que la exposición a neumoalergenos en edades tempranas de la vida es un factor de riesgo para el desarrollo de asma. Sporik et[3].

El tabaco aumenta inespecíficamente la reactividad bronquial, puede que por aumento de la inflamación bronquial. El que una madre sea fumadora, aumenta el riesgo de comienzo de asma y de exacerbaciones de asma. La exposición intrauterina al humo de tabaco puede afectar la reactividad bronquial y producir una alteración inicial de la función pulmonar al nacimiento[4]

Está demostrado que la contaminación ambiental, por ejemplo por dióxido de nitrógeno y ozono, puede aumentar el efecto de los alergenos, posiblemente porque se incrementa la reactividad bronquial.

2.2 OBJETIVOS DEL ESTUDIO

[3] Sporik R, Holgate S, Platts-Mills T, Cogswell J. Exposure to house-dust mite allergen (Der p I) and the development of asthma in childhood. N Engl Med J 1990; 23: 502-507.
[4] Hanrahan JP.; Tager IB.; Segal MR.; Tosteson TD.; Castile RG.; van Vunakis H. et al. The effect of maternal smoking during pregnancy on early infant lung function. Am Rev Respir Dis 1992; 145: 1129-35.

Objetivo General

Aplicación de la Regresión Logística a un caso del área de la epidemiología: el asma en niños escolares de tres a catorce años en la ciudad de Moquegua.

Conocer la prevalencia, así como los factores asociados a la presencia de asma en escolares de 3 a 14 años de edad de los centros educativos nacionales y particulares en la ciudad de Moquegua.

Objetivos Específicos

- Determinar la prevalencia de asma en escolares de 3 a 14 años de la ciudad de Moquegua.
- Identificar los factores asociados a la presencia de asma en los escolares de 3 a 14 años.

Conocer la prevalencia del asma es importante para que las organizaciones de salud puedan adoptar medidas correctivas, como la implementación de programas de prevención, ya que una elevada prevalencia y el carácter crónico del asma ocasionan que ésta resulte una enfermedad “cara” para la sociedad.

A continuación se presentan algunas definiciones necesarias para una mejor comprensión del problema:

2.3. CONCEPTOS BÁSICOS

1. EPIDEMIOLOGÍA

La epidemiología tiene entre uno de sus objetivos primordiales el estudio de la distribución y los determinantes de las diferentes enfermedades. La cuantificación y la medida de la enfermedad o de otras variables de interés son elementos fundamentales para formular y testar hipótesis, así como para permitir comparar las frecuencias de enfermedad entre diferentes poblaciones o entre personas con o sin una exposición o característica dentro de una población determinada.

2. PREVALENCIA

Cuantifica la proporción de individuos de una población que presentan el evento de interés en un momento o periodo de tiempo determinado. Su cálculo se estima mediante la expresión:

$$P = \frac{N^{o}\,eventos}{N^{o}\,individuosdelapoblación}$$

Indica la "carga" del evento que soporta la población, su valor oscila entre 0 y 1, aunque a veces se expresa como porcentaje.

3. HACINAMIENTO

Se considera como hogares con hacinamiento a aquellos con más de tres personas por habitación, excluyendo el baño, cocina pasadizos y garaje.

4. DERMATITIS ATÓPICA

La dermatitis atópica es una irritación en la piel que, puede ser, de corta o larga duración. También se conoce con el nombre de eczema. Aunque la dermatitis atópica es más común en los niños, cualquier persona, puede presentar este problema en la piel. Se piensa que los lactantes que la presentan tienden a desarrollar posteriormente asma.

5. ETIOPATOGENIA

Término formado a su vez por otros dos. Etiología, que hace referencia a la causa o causas de una enfermedad además de factores propios del paciente que la favorecerían y factores propios de la enfermedad. La Patogenia serían los mecanismos por los cuales se desencadena la enfermedad.

2.4. DISEÑO DE LA INVESTIGACIÓN

El presente es un estudio de corte transversal, desarrollado en base a los resultados de la “Encuesta Sobre Enfermedades Respiratorias 2004” efectuado por el Ministerio de Salud.

2.5. DISEÑO MUESTRAL

La población objetivo **estuvo conformada por escolares de 3 a 14 años de edad, que asisten de manera regular a centros educativos tanto nacionales como particulares de la ciudad de Moquegua.**
La selección de la muestra **fue aleatoria, bajo un diseño muestral probabilístico elaborado por el Ministerio de Salud. Se seleccionaron 27 instituciones educativas, en ellos se entrevistó a 959 escolares, de los cuales 716 (74.7%)**

pertenecen a centros educativos nacionales y 243 (25.3%) a centros educativos particulares.

El tipo de entrevista fue directa, se aplicó a los escolares seleccionados, así como a sus padres.

El instrumento de medición: cuestionario, para este estudio consta de las siguientes secciones:

1 Datos generales del alumno

2 Información del cuadro clínico

3 Antecedentes personales y familiares

4 Información ambiental de la ciudad de Moquegua

2.6 DEFINICIÓN OPERACIONAL DE VARIABLES

La definición operacional de las variables que se consideraron en esta investigación son las siguientes:

1. Riesgo por sexo

Según la literatura médica, el sexo masculino se considera como riesgo para la presencia de asma.

2. Riesgo por grupo de edad

Entre los grupos de edad formados para el estudio: [3,9] y [10,14] en años cumplidos, se considera como factor de riesgo el que el escolar pertenezca al grupo de [3,9] años.

3. Riesgo por presencia de gatos en el hogar

Debido a que el asma es una enfermedad alérgica, la presencia de gatos es un factor de riesgo en la aparición de dicha enfermedad.

4. Riesgo por presentar dermatitis atópica

Se preguntó acerca de si el escolar presentó alguna vez un cuadro de dermatitis atópica.

5. Riesgo por fumadores en casa

Se determinó en base a la información proporcionada por los padres del escolar de acuerdo al número de cigarros que se fuma diariamente en su casa. Se consideró como riesgo en los hogares donde se fumaba habitualmente.

6. Riesgo por fumar durante el embarazo

Se determinó en base a la frecuencia y al número de cigarros que fumó la madre del escolar durante el período de gestación.

7. Riesgo por historia familiar de alergia

Se preguntó a los padres del escolar, si presentaban algún tipo de alergia.

8. Riesgo por no recibir lactancia materna

La madre de familia fue interrogada acerca de si su hijo había recibido lactancia materna, se considera como riesgo el que el escolar no haya recibido lactancia materna.

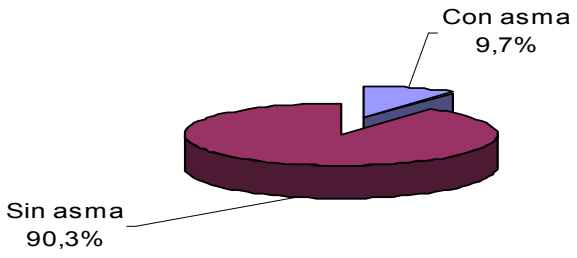
2.7 ANÁLISIS ESTADÍSTICO

El análisis estadístico se basó en un análisis descriptivo de las variables socio-demográficas y otras involucradas en el estudio. Asimismo se aplicó el análisis de regresión logística múltiple para estudiar la asociación entre la condición de asma y las variables predictoras.

2.7.1 ANÁLISIS DESCRIPTIVO DE DATOS

La prevalencia de asma en la muestra es de 9.7% (93), es decir, 9,7 de cada 100 escolares entre los 3 y 14 años de la ciudad de Moquegua presenta asma.

GRÀFICO Nº 1: PREVALENCIA DE ASMA



En el cuadro Nº 1 se muestra los resultados para las variables socio-demográficas por presencia de asma y a nivel total.

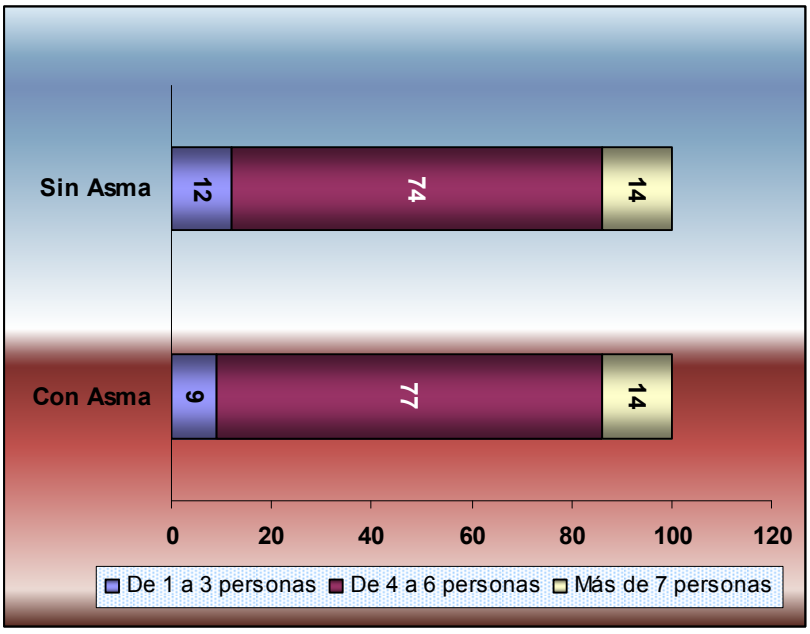
CUADRO Nº 1 VARIABLES SOCIO - DEMOGRÁFICAS POR PRESENCIA DE ASMA

		Asma		
		Sin Asma	Con Asma	Total
TOTAL ESCOLARES ENTREVISTADOS		**866**	**93**	**959**
Sexo	Masculino	50.6%	57.0%	51.2%
	Femenino	49.4%	43.0%	48.8%
Grupo de edad	De 3 a 9 años	57.6%	65.6%	58.4%
	De 10 a 14 años	42.4%	34.4%	41.6%
Tipo de colegio	Estatal	73.6%	84.9%	74.7%
	Particular	26.4%	15.1%	25.3%
Estrato de la vivienda	Estrato Bajo	44.6%	44.1%	44.5%
	Estrato Medio	45.7%	43.0%	45.5%
	Estrato Alto	9.7%	12.9%	10.0%
Numero de miembros del hogar	De 1 a 3 personas	12.0%	8.6%	11.7%
	De 4 a 6 personas	73.6%	77.4%	73.9%
	Más de 7 personas	14.4%	14.0%	14.4%
Hacinamiento	Sin hacinamiento	93.8%	95.7%	94.0%
	Con hacinamiento	6.2%	4.3%	6.0%

La distribución por sexo en los escolares es similar, sólo en los que presentan cuadro de asma la proporción de hombres es ligeramente mayor (57%).
La mayoría de escolares entrevistados pertenece al grupo de edad de 3 a 9 años (58.4%), la proporción de aquellos que asisten a colegios particulares es casi tres veces (74.7%) de los que asisten a colegios del estado (25.3%). El porcentaje de escolares cuyas viviendas pertenecen al estrato alto y medio es similar; sólo el 6% de los hogares entrevistados tiene hacinamiento.

La mayoría de hogares de los escolares entrevistados tienen entre 4 a 6 miembros (73.9%), exploratoriamente se puede observar que existe mínima diferencia en esta variable por presencia de asma como se aprecia en el siguiente gráfico.

GRÀFICO Nº 2: INTEGRANTES DEL HOGAR POR PRESENCIA DE ASMA



En el cuadro Nº 2, se aprecia que el riesgo de los escolares por pertenecer al sexo masculino es del 51,2% a nivel global, siendo similar su distribución por presencia de asma. La mayoría de escolares entrevistados recibió lactancia materna, por ello el riesgo por no recibir lactancia materna es apenas de 5,3%, sin embargo esta proporción es menor en aquellos que no tienen asma comparado con los que la tienen.

Algo que es importante resaltar es la proporción de escolares que recibieron lactancia materna por un período mayor a 6 meses, en forma global este porcentaje es mayor al de aquellos que recibieron lactancia por menos de 6 meses, sin embargo al hacer el análisis por presencia de asma vemos que en los escolares que presentaron asma el 96.8% sólo recibió lactancia materna por un período menor a 6 meses, este hecho nos hace sospechar que se trata de uno de los factores de riesgo más importantes en la presencia de asma.

Observamos que la mayoría de las madres de los escolares entrevistados no fumó durante el período de gestación (97.7%), además el 82.9% de los padres de familia no presentó antecedentes familiares de alergia.

CUADRO Nº 2 VARIABLES ASOCIADAS A PRESENCIA DE ASMA

		Asma		
		Sin Asma	Con Asma	Total
Riesgo por el sexo	Femenino	49,4%	43,0%	48,8%
	Masculino	50,6%	57,0%	51,2%
Total		100,0%	100,0%	100,0%
Riesgo por el grupo de edad	10 - 14 años	55,0%	58,1%	55,3%
	3 - 9 años	45,0%	41,9%	44,7%
Total		100,0%	100,0%	100,0%
Riesgo por presencia de gatos en la vivienda	No	92,0%	93,5%	92,2%
	Si	8,0%	6,5%	7,8%
Total		100,0%	100,0%	100,0%
Riesgo por dermatitis atopica	No	83,7%	75,3%	82,9%
	Si	16,3%	24,7%	17,1%
Total		100,0%	100,0%	100,0%
Riesgo por fumadores en casa	No	85,1%	79,6%	84,6%
	Si	14,9%	20,4%	15,4%
Total		100,0%	100,0%	100,0%
Riesgo por fumar durante embarazo	No	98,3%	92,5%	97,7%
	Si	1,7%	7,5%	2,3%
Total		100,0%	100,0%	100,0%
Riesgo por historia familiar de alergia	No	83,4%	78,5%	82,9%
	Si	16,6%	21,5%	17,1%
Total		100,0%	100,0%	100,0%
Riesgo por no lactancia materna	No	95,2%	90,3%	94,7%
	Sí	4,8%	9,7%	5,3%
Total		100,0%	100,0%	100,0%
Tiempo de lactancia materna	menos de 6 meses	4,8%	96,8%	13,8%
	más de 6 meses	95,2%	3,2%	86,2%
Total de Escolares		**866**	**93**	**959**
Total %		100,0%	100,0%	100,0%

En el 15,4% de los hogares de los escolares, existieron hábitos de fumar (Riego por fumadores en casa), sin embargo para los escolares con presencia de asma este porcentaje es mayor (20,4%) comparado con el grupo de escolares que no presenta asma; la dermatitis atópica sólo se presentó en el 17.1% de escolares entrevistados, siendo mayor para aquellos escolares con asma (24.7%). La presencia de gatos en los hogares de los escolares entrevistados es apenas de 7.8%.

Al realizar el análisis de cada uno de los posibles factores asociados a la presencia de asma, observamos que la variable Tiempo de lactancia materna presenta sólo 3 observaciones para aquellos escolares con asma que lactaron por más de 6 meses (Cuadro Nº 1 del Apéndice), este hecho influye en la estimación de parámetros, lo cual haría que las estimaciones no sean confiables, por ello no será considerado en el análisis de regresión logística.

En el siguiente cuadro podemos apreciar que de todos los posibles factores estudiados, solamente resultan significativamente asociados a la presencia de asma las variables: Riesgo por dermatitis atópica y Riesgo por fumar durante el embarazo.

CUADRO Nº 3 ODDS RATIO

VARIABLES	OR	I.C. AL 95%	
		L.I	L.S
Riesgo por sexo	1,29	0,84	1,99
Riesgo por grupo de edad	0,88	0,57	1,36
Riesgo por presencia de gatos	0,80	0,34	1,89
Riesgo por dermatitis atópica	1,69	1,02	2,80
Riesgo por fumadores en casa	1,47	0,86	2,51
Riego por fumar durante el embarazo	4,62	1,83	11,64
Riesgo por historia alérgica familiar	1,37	0,81	2,32
Riesgo por no lactancia materna	2,10	0,99	4,47

2.7.2 ANÁLISIS DE REGRESIÓN LOGÍSTICA

A través del análisis de regresión logística determinaremos los factores asociados a la presencia de asma en escolares de 3 a 14 años de la ciudad de Moquegua (Y).

Proponemos como posibles factores a las siguientes variables:

Riesgo por sexo (X_1), Riesgo por grupo de edad (X_2), Riesgo por presencia de gatos (X_3), Riesgo por dermatitis atópica (X_4), Riesgo por fumadores en casa (X_5),

Riesgo por fumar durante el embarazo (X_6), Riesgo por historia familiar de alergia (X_7), Riesgo por no lactancia materna (X_8).

El Modelo de Regresión Logística inicial que se plantea tiene la siguiente forma:

$$E(Y/X_1,X_2,...X_8)=P(Y=1/X)=\frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_8X_8)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_8X_8)}$$

Donde:

P = Probabilidad de presencia de asma y X_1, X_2,....X_8 son las variables predictoras.

Con el método de selección completo se procedió a elegir las variables explicativas para la presencia de asma en los escolares.

La hipótesis planteada con el análisis de regresión logística es:

H_0 : $\beta_1=\beta_2=\beta_3=....=\beta_8=0$ Ninguna variable explica la presencia de asma

H_1 : Al menos un $\beta_j\neq 0$ j=1,2,….,8

A través del análisis de regresión logística, puede observarse que la mejor variable predictora para separar el grupo de asmáticos del grupo de no asmáticos es el Riesgo por fumar durante el embarazo (X_6), sin embargo se consideran adicionalmente las variables Riesgo por no lactancia materna (X8), Riesgo por dermatitis atópica (X2), y Riesgo por historia familiar de alergia (X7) por cuanto según la literatura estudiada constituyen variables que influyen en la presencia de asma.

Los resultados de las estimaciones de los parámetros del modelo se exponen en el siguiente cuadro.

CUADRO Nº 4 ESTIMACIÓN DE PARÁMETROS

	B	Wald	p-valor	Exp(B)	I.C.95,0% para EXP(B)	
					inferior	superior
Riesgo por dermatitis atópica	0,393	2,206	0,137	1,482	0,882	2,490
Riesgo por fumar durante el embarazo	1,356	7,831	0,005	3,882	1,501	10,036
Riego por historia alérgica familiar	0,294	1,160	0,281	1,341	0,786	2,288
Riesgo por no lactancia materna	0,586	2,160	0,142	1,796	0,822	3,924
Constante	-2,460	306,432	0,000	0,085		

De este modo, el modelo final ajustado es:

$$E(Y/X_1,X_2,...,X_8)=\frac{\exp(-2.46+0.393X_4+1.356X_6+0.294X_7+0.586X_8)}{1+\exp(-2.46+0.393X_4+1.356X_6+0.294X_7+0.586X_8)}$$

Evaluamos la bondad de ajuste de nuestro modelo con la prueba ómnibus (Estadística Chi cuadrado) y la prueba de Hosmer y Lemeshow, las cuales resultaron significativas con p=0.007 y p=0.09, respectivamente (Cuadro Nº 2 y 3 del Apéndice) evidenciando el buen ajuste del modelo.

Por lo tanto, se rechaza la hipótesis nula que todos los parámetros asociados a las variables consideradas en el modelo son nulos, lo cual quiere decir que al menos una de las variables contribuye a explicar la presencia de asma en los escolares.

El estadístico de Wald nos ayuda a determinar cual o cuales son las variables cuyos parámetros asociados son iguales a cero, para lo cual se plantean las siguientes hipótesis:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

En el Cuadro Nº 4, observamos que la única variable que resultó estadísticamente significativa fue Riesgo por fumar durante el embarazo. El motivo por el cual las otras variables no resultaron significativas en este estudio podría ser la baja frecuencia que estas presentan en algunas categorías (Cuadro Nº 1.9 del Apéndice)

Interpretación de parámetros

Con un nivel de significación del 5% podemos concluir que el parámetro asociado a la variable Riesgo por fumar durante el embarazo es estadísticamente distinto de cero, es decir, significativo (p-valor igual a 0,005).

Existe 3,88 veces más posibilidad que un escolar cuya madre fumó durante el embarazo presente asma comparado con el escolar cuya madre no fumó durante su embarazo.

Existe 1,48 veces más chance que un escolar que padeció dermatitis atópica presente asma que el que no la padeció.

La chance de presentar asma cuando se tiene antecedentes de historia alérgica familiar es de 1,34 veces más comparado con aquel que no tiene historia alérgica familiar.

Existe 1,79 veces más chance de presentar asma, si el escolar no recibió lactancia materna comparado con el que si la recibió.

Capacidad Predictiva del modelo

Como se mencionó anteriormente se incluyó en el modelo las variables: Riesgo por no lactancia materna (X8), Riesgo por dermatitis atópica (X2), y Riesgo por historia familiar de alergia (X7), a pesar que no resultaron significativas, debido a que mejoran la capacidad predictiva del modelo. Sí sólo consideramos el Riesgo por fumar durante el embarazo en el modelo, el máximo valor de sensibiidad que obtenemos es del 7,5% mientras que incluyendo las variables: alcanzamos como valor máximo el 48.4%

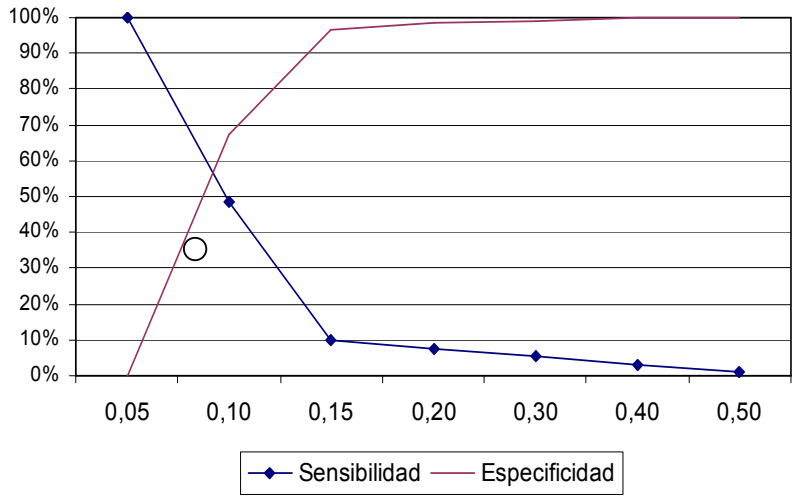
Una herramienta que nos ayuda a medir la bondad de ajuste del modelo es la tabla de clasificación.

CUADRO Nº 4: Sensibilidad y Especificidad

Punto de Corte	Sensibilidad	Especificidad	Global
0,05	100,0%	0,0%	9,7%
0,10	48,4%	67,2%	65,4%
0,15	9,7%	96,3%	87,9%
0,20	7,5%	98,3%	89,5%
0,30	5,4%	99,2%	90,1%
0,40	3,2%	100,0%	90,6%
0,50	1,1%	100,0%	90,4%

Para fijar el punto de corte adecuado a partir del cual se considera la presencia de asma en los escolares, graficamos los valores de sensibilidad y especificidad que calcula el modelo para distintos niveles de probabilidad, donde se cruzan ambas curvas encontramos el punto que maximiza ambos indicadores.

GRÀFICO Nº 2: Sensibilidad y Especificidad



Como observamos en el gráfico anterior, el punto de corte más adecuado es 0.1; la probabilidad de clasificar de forma correcta a los escolares en cada grupo siguiendo

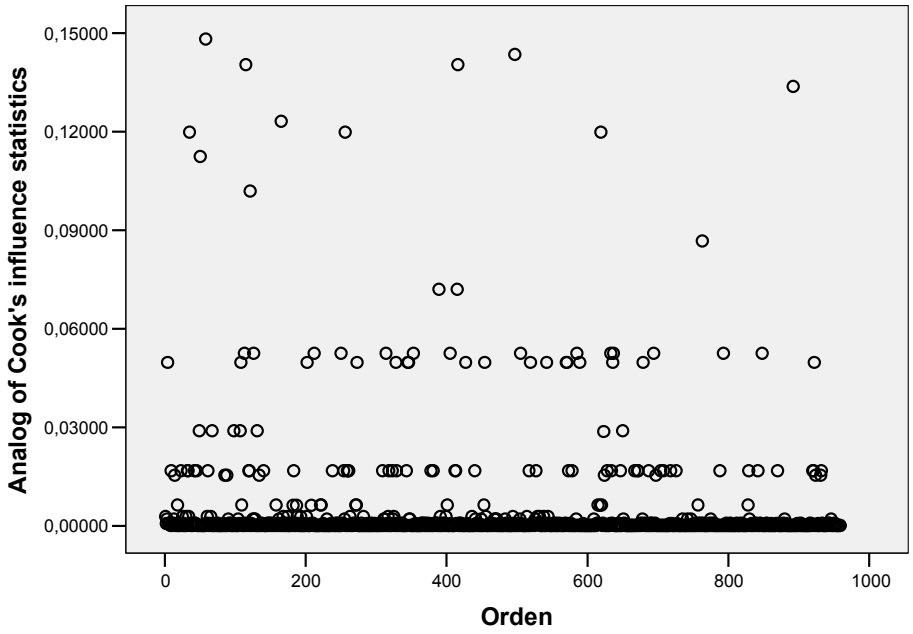
este modelo fue del 65,4%. La sensibilidad obtenida fue de 48.4%, es decir, el 48.4% de escolares con asma fueron clasificados correctamente por el modelo ajustado. Asimismo la especificidad de 67.2% nos dice que el 67.2% de los escolares que no presentaron asma fueron clasificados correctamente. Estos resultados confirman el buen ajuste del modelo, aunque su potencial como modelo predictivo es bajo.

El motivo por el cual otras variables importantes asociadas a la presencia de asma según la literatura médica no resultan significativas en el estudio, sería las bajas frecuencias que presentan, debido a las características propias de la zona.

Diagnóstico del Modelo

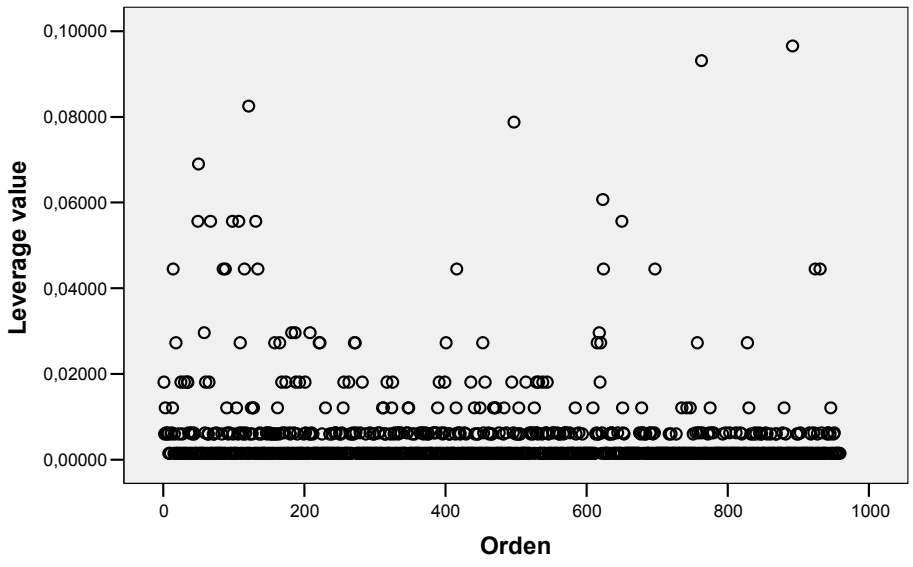
Para evaluar la adecuacidad de nuestro modelo procedemos a realizar el análisis de influencia. Este análisis mide la influencia de las observaciones en la estimación de los parámetros y probabilidad del modelo, cuanto más grande sea el valor se dice que dicha observación ejerce mayor influencia en la estimación del modelo.
El gráfico Nº 3 muestra los valores para la Distancia de Cook, como se puede apreciar estos son valores pequeños, menores que uno. Por lo tanto no hay presencia de datos influyentes.

GRÀFICO Nº 3: ESTADÍSTICO DE COOK VS. ORDEN



Al graficar las medidas de apalancamiento (Leverage), los cuales detectan las observaciones que ejercen influencia en los valores predichos, observamos que no existen datos potencialmente influyentes.

GRÀFICO Nº 4: VALOR DE INFLUENCIA VS. ORDEN



CONCLUSIONES Y RECOMENDACIONES

- ✓ En nuestro estudio, el análisis descriptivo permitió determinar que la prevalencia de asma en los escolares de la ciudad de Moquegua es de 9.7%, es decir, 9.7 de cada 100 escolares de 3 a 14 años de la ciudad de Moquegua padece de asma.

- ✓ De acuerdo a los resultados del análisis de regresión logística, el Riesgo por fumar durante el embarazo constituye el factor de riesgo más importante para la presencia de asma en la población de estudio.

- ✓ La posibilidad de que un escolar presente asma si su madre fumó durante el embarazo es 3.88 veces mayor comparado con el escolar cuya madre no fumó durante el embarazo.

- ✓ El riesgo de padecer asma para un escolar que tuvo antecedentes familiares de historia alérgica es 1.34 veces mayor con respecto al escolar que no tuvo antecedentes familiares de alergia.

- ✓ Existe 1.48 veces más chance que un escolar que tuvo dermatitis atópica padezca de asma comparado con el escolar que no la tuvo.

- ✓ EL modelo permitió que el 48.4% de los escolares que padecen asma sean correctamente clasificados por el modelo, y el 67.2% de los escolares que no

padecen de asma fueron también correctamente clasificados por el modelo. A nivel global, el 65.4% de los escolares fueron correctamente clasificados por el modelo

- ✓ La causa por la cual las variables consideradas como factores importantes para la presencia de asma no resultan significativas en el presente estudio, sería la baja frecuencia con que se manifiestan en esta población en particular debido a las características particulares de la zona.

BIBLIOGRAFÍA

1.Gilberto A. Paula. (2004) Modelos de Regressao com apoio computacional

2. Hosmer, D. Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons. New York, USA

3. Kleinbaum, David G. (1998). Logistic Regression A Self – Learning Text. Springer- Verlag New York. Inc.

4. Montgomery, Douglas (2002). Introducción al análisis de regresión Lineal. Editorial Continental

5. Agresti, Alan (1996). An Introduction to categorical data análisis. John Wiley & Sons. New York, USA

6. Mc. Cullagh y Nelder (1991). Generalized Linear Models. Second edition Chapman Ktall

7. Visauta Vinacua, B. (1998). Análisis estadístico con SPSS para windows. Estadística multivariante. McGraw – Hill

8. Asma Bronquial. Página electrónica:

http://www.tuotromedico.com/temas/asma.htm

9. Tras las huellas del Asma. Página electrónica:

http://www.elmundo.es/salud/1997/250/01888.html

10. Boletín Electrónico del hospital dos de mayo. Página electrónica:

http://www.respirar.org/forolatino/peru.htm

APÉNDICE

CUADRO Nº 1: TABLAS DE CONTINGENCIA

1.1 Riesgo por el sexo * Asma

Count

		Asma		
		No	Si	Total
Riesgo por el sexo	Femenino	428	40	468
	Masculino	438	53	491
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por el sexo (Femenino / Masculino)	1,295	,841	1,994
For cohort Asma = No	1,025	,984	1,069
For cohort Asma = Si	,792	,536	1,170
N of Valid Cases	959		

1.2 Riesgo por el grupo de edad * Asma

Count

		Asma		
		No	Si	Total
Riesgo por el grupo de edad	10 - 14 años	476	54	530
	3 - 9 años	390	39	429
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por el grupo de edad (10 - 14 años / 3 - 9 años)	,881	,572	1,359
For cohort Asma = No	,988	,948	1,030
For cohort Asma = Si	1,121	,758	1,658
N of Valid Cases	959		

1.3 Riesgo por presencia de gatos en la vivienda * Asma

Count

		Asma		
		No	Si	Total
Riesgo por presencia de gatos en la vivienda	No	797	87	884
	Si	69	6	75
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por presencia de gatos en la vivienda (No / Si)	,797	,336	1,888
For cohort Asma = No	,980	,914	1,051
For cohort Asma = Si	1,230	,557	2,719
N of Valid Cases	959		

1.4 Riesgo por dermatitis atopica * Asma

Count

		Asma		
		No	Si	Total
Riesgo por dermatitis atopica	No	725	70	795
	Si	141	23	164
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por dermatitis atopica (No / Si)	1,689	1,020	2,798
For cohort Asma = No	1,061	,993	1,132
For cohort Asma = Si	,628	,404	,975
N of Valid Cases	959		

1.5 Riesgo por fumadores en casa * Asma

Count

		Asma		
		No	Si	Total
Riesgo por fumadores en casa	No	737	74	811
	Si	129	19	148
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por fumadores en casa (No / Si)	1,467	,857	2,511
For cohort Asma = No	1,043	,976	1,113
For cohort Asma = Si	,711	,443	1,140
N of Valid Cases	959		

1.6 Riesgo por fumar durante embarazo * Asma

Count

		Asma		
		No	Si	Total
Riesgo por fumar durante embarazo	No	851	86	937
	Si	15	7	22
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por fumar durante embarazo (No / Si)	4,618	1,833	11,636
For cohort Asma = No	1,332	1,001	1,773
For cohort Asma = Si	,288	,151	,549
N of Valid Cases	959		

1.7 Riesgo por historia familiar de alergia * Asma

Count

		Asma		
		No	Si	Total
Riesgo por historia familiar de alergia	No	722	73	795
	Si	144	20	164
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por historia familiar de alergia (No / Si)	1,374	,812	2,324
For cohort Asma = No	1,034	,973	1,100
For cohort Asma = Si	,753	,473	1,199
N of Valid Cases	959		

1.8 Riesgo por no lactancia materna * Asma

Count

		Asma		
		No	Si	Total
Riesgo por no lactancia materna	No	824	84	908
	Sí	42	9	51
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por no lactancia materna (No / Sí)	2,102	,989	4,468
For cohort Asma = No	1,102	,969	1,253
For cohort Asma = Si	,524	,280	,981
N of Valid Cases	959		

1.9 Riesgo por tiempo de lactancia materna * Asma

Crosstab

Count

		Asma		
		No	Si	Total
Riesgo por tiempo de lactancia materna	,00	824	3	827
	1,00	42	90	132
Total		866	93	959

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Riesgo por tiempo de lactancia materna (,00 / 1,00)	588,571	178,816	1937,279
For cohort Asma = No	3,131	2,439	4,020
For cohort Asma = Si	,005	,002	,017
N of Valid Cases	959		

CUADRO Nº 2: PRUEBA ÓMNIBUS DE BONDAD DE AJUSTE

Prueba omnibus sobre los coeficientes del modelo

		Chi-square	df	Sig.
Step 1	Step	14,218	4	,007
	Block	14,218	4	,007
	Model	14,218	4	,007

CUADRO Nº 3: PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

Prueba de Hosmer y Lemeshow

Step	Chi-square	df	Sig.
1	4,766	2	,092

Variables en el modelo

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								inferior	superior
a	rderato(1)	,393	,265	2,206	1	,137	1,482	,882	2,490
	rfumemb(1)	1,356	,485	7,831	1	,005	3,882	1,501	10,036
	rhisale(1)	,294	,273	1,160	1	,281	1,341	,786	2,288
	rnlactmat(1)	,586	,399	2,160	1	,142	1,796	,822	3,924
	Constant	-2,460	,141	306,432	1	,000	,085		

a. Variable(s) entered on step 1: rderato, rfumemb, rhisale, rnlactmat.

CUADRO Nº 4: ESTIMACIÓN DE PARÁMETROS

CUADRO Nº 5: TABLA DE CLASIFICACIÓN PARA PRESENCIA DE ASMA

Tabla de Clasificación[a]

		Pronósticado		
		Asma		Porcentaje
Observado		No	Si	Correcto
Asma	No	582	284	67,2
	Si	48	45	48,4
Porcentage global				65,4

a. El punto de corte es 0 ,10